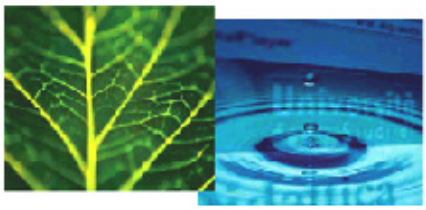
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Analysis of the PC algorithm as a tool for the inference of gene regulatory networks: evaluation of the performance, modification and application to selected case studies.

Emanuela Coller

Advisor:

Prof. Enrico Blanzieri

Università degli Studi di Trento

Co-advisor:

Dott. Claudio Moser

Fondazione Edmund Mach

April 2013

A Tulio ed Annamaria

Il giorno piú bello? . . . Oggi.
La cosa piú facile? . . . Sbagliarsi.
L'ostacolo piú grande? . . . La paura.
Lo sbaglio peggiore? . . . Arrendersi.
La radice di tutti i mali? . . . L' egoismo.
La distrazione piú bella? . . . Il lavoro.
La peggiore sconfitta? . . . Lo scoraggiamento.
I migliori insegnanti?. . . I bambini.
La prima necessitá?. . . Parlare con gli altri.
La cosa che piú fa felici? . . . Essere di aiuto agli altri.
Il Mistero piú grande? . . . La morte.
Il peggiore difetto? . . . Il malumore.
La persona piú pericolosa?. . . Il bugiardo.
Il sentimento piú dannoso?. . . Il rancore.
Il regalo piú bello? . . . Il perdono.
La cosa di cui non se ne puó fare a meno? . . . La casa.
La strada piú rapida? . . . Il cammino giusto.
La sensazione piú gratificante? . . . La pace interiore.
Il gesto piú efficace? . . . Il sorriso.
Il migliore rimedio? . . . L' ottimismo.
La maggiore soddisfazione? . . . Il dovere compiuto.
La forza piú potente del mondo? . . . La fede.
Le persone piú necessarie? . . . I genitori.
La cosa piú bella di tutte? . . . L' AMORE!

Maria Teresa di Calcutta

Abstract

The expansion of a Gene Regulatory Network (GRN) by finding additional causally-related genes, is of great importance for our knowledge of biological systems and therefore relevant for its biomedical and biotechnological applications.

Aim of the thesis work is the development and evaluation of a bioinformatic method for GRN expansion. The method, named PC-IM, is based on the PC algorithm that discovers causal relationships starting from purely observational data. PC-IM adopts an iterative approach that overcomes the limitations of previous applications of PC to GRN discovery.

PC-IM takes in input the prior knowledge of a GRN (represented by nodes and relationships) and gene expression data. The output is a list of genes which expands the known GRN. Each gene in the list is ranked depending on the frequency it appears causally relevant, normalized to the number of times it was possible to find it. Since each frequency value is associated with precision and sensitivity values calculated using the prior knowledge of the GRN, the method provides in output those genes that are above the value of frequency that optimize precision and sensitivity (cut-off frequency).

In order to investigate the characteristics and the performances of PC-IM, in this thesis work several parameters have been evaluated such as the influence of the type and size of input gene expression data, of the number of iterations and of the type of GRN. A comparative analysis of PC-IM versus another recent expansion method (GENIES) has been also performed.

Finally, PC-IM has been applied to expand two real GRNs of the model plant Arabidopsis thaliana.

Keywords[bioinformatics, iterative method, PC algorithm, expansion, causal relationship, gene regulatory network, FOS-GRN]

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The genome is the entire genetic material (DNA or RNA in many types of virus) of an organism (both unicellular and multicellular). It plays a central role in the control of all cellular processes (e.g. the response of a cell to environmental signals, the differentiation of cells and groups of cells in the unfolding of developmental programs, the replication of the DNA preceding cell division). The central dogma of molecular biology says that the genetic material (DNA) is transcribed into RNA (transcription process) and then translated into protein (translation process). This is the basic mechanism of gene expression and it relies upon a unidirectional flow of the genetic information. Gene expression is finely regulated within the cell [Lewin and Dover, 1994] both at transcription and translation levels and this control is essential to maintain cell homeostasis and to allow the organism life. Proteins may function as:

- transcription factors binding to regulatory sites of other genes;
- enzymes catalyzing metabolic reactions;
- structural components of the cell;
- components of signal transduction pathways.

Different proteins may regulate the same gene or may form a single gene regulatory complex. Two genes can have a causal interaction without having a physical interaction. In fact there are indirect regulations via proteins and metabolism [Lauria and di Bernardo, 2010]. This variety of phenomena that regulates gene expression can be represented by Gene Regulatory Network (GRN).

GRNs are the complex systems that are formed from the regulatory interactions (causal relationships) between DNA, RNA and proteins. The final expression of a gene is determined from these regulatory interactions between genes and proteins. In a biological cell

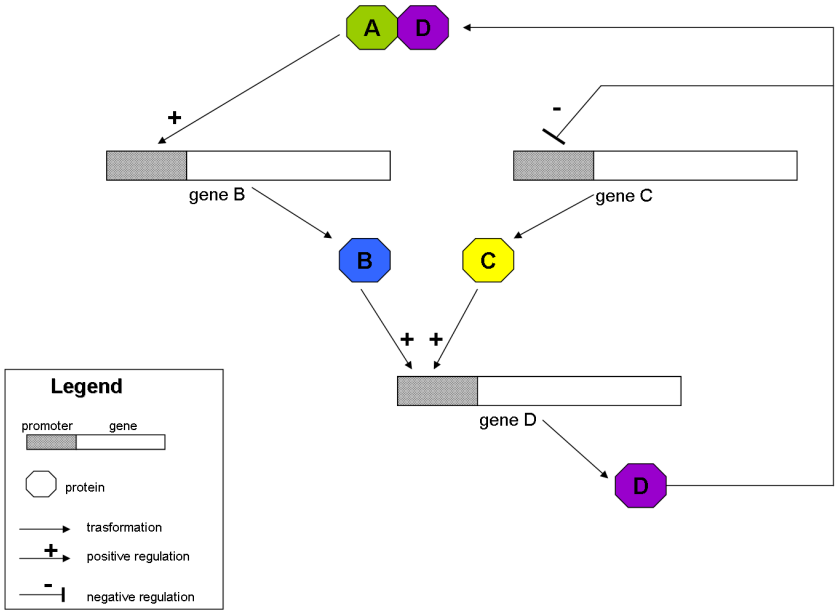


Figure 1.1: **Example of a gene regulatory network.**
Protein B and C independently activate gene D by binding to different regulatory sites on the promoter of gene D. Protein D represses gene C and interacts with protein A to activate gene B.

there are positive and negative regulations. In the positive regulation (or activation) the regulator activates the target genes, instead in the negative regulation (or inhibition) the regulator inhibits the target genes. Figure 1.1 reports an example of the gene regulatory network. More complex graphical conventions to represent cellular networks are proposed by Kohn [1999] and Kohn et al. [2006].

One of the objectives of molecular biology is to understand the regulatory mechanisms behind biological processes. This implies that a full description of a GRN determines the identification of the genes comprised in it, the comprehension of the gene connections (functional relations) and the elucidation of the kind of relationships between the genes of the GRN. A correct description of a GRN is of the greatest importance since it will allow either predicting the behavior of the system under perturbation or manipulating it for a specific aim [Bansal et al., 2007]. The problem is that the knowledge of biological systems is incomplete, therefore the construction of putative biological models and GRNs are necessarily based on incomplete information.

An approach to this problem is to adopt the principles of reverse engineering. Reverse engineering is the process that, starting from iterative experimentation (for example gene expression data) on an unknown system, arrives to the reconstruction of GRNs. Figure 1.2 is a schematic drawing of the process of reverse engineering. The strategy starts from

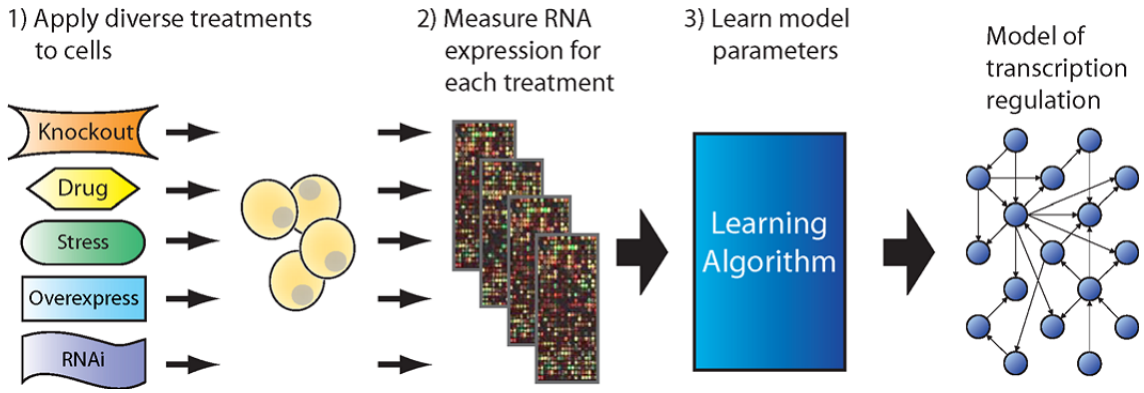


Figure 1.2: **General reverse engineering to infer GRNs.**
(taken from [Gardner and Faith, 2005]).

experiments of cell perturbation with various treatments. In the second step, the aim is to measure the expression of the transcripts. Subsequently, a learning algorithm infers the model of transcription regulation using the expression data. The final result is the gene regulatory network.

This approach requires large data sets and extensive computational resources, because there is a big number of network architectures that are compatible with the same experiment results (set of expression data) [Tegner et al., 2003]. Luckily, in recent times the quantity of information that is available for reverse engineering is enormous. In fact, the genome projects have rapidly generated large datasets of sequences of genes and proteins that govern cellular behavior. Moreover 20 years ago [Schena et al., 1996] [Chee et al., 1996] [Lockhart et al., 1996] gene expression microarrays permitting of simultaneously measure thousands of transcripts [Schwarz, 1978]. The array technology has several limitations [Marioni et al., 2008]:

- background levels of hybridization limit the accuracy of expression measurements, particularly for transcripts present in low abundance;
- probes differ considerably in their hybridization properties and this affects the comparison of hybridization results across arrays;
- arrays are limited to measure abundance of transcripts with relevant probes on the array;
- arrays do not allow to measure DNA methylation and other DNA modifications.

Sequencing-based approaches to measure gene expression levels have the potential to overcome these limitations (454 Life Sciences -Roche- [Margulies et al., 2005], Illumina-Solexa sequencing- [Bennett et al., 2005]). Despite this, microarrays are still widely used,

because the new technologies are complex and data available on public sites are still limited both in quantity and in number of different organisms analysed.

1.1 Objective of the Thesis

The aim of this thesis work is the development of a method to expand a characterized Gene Regulatory Network, called Local Gene Network (LGN) in the following. The expansion leads to the identification of new genes that are related with the known genes of the LGN. These are listed in a final expansion gene list which reports as well an estimate of their reliability. The expansion genes are obtained analyzing all the genes of interest given in the input list together with the corresponding gene expression data loaded by the user. Before expanding a LGN, the whole input gene list is subdivided in subgroups (called *tiles*). Each *tile* contains different genes of the input gene set, but all include the genes of the LGN. The whole expansion procedure is iterated i times and the final output takes into account the output of all the iterations. Subsequently, the reliability is determined by an intrinsic performance evaluation. This step requires to calculate precision and sensitivity within the genes of the LGN and then to project these values on the new genes. The procedure is called PC-Iterative Method (PC-IM). The term PC indicates that our algorithm uses the causal discovery algorithm developed by Spirtes and Glymour (PC algorithm) [Spirtes and Glymour, 1991], instead the "iterative"term indicates that the analysis of the whole input gene-set is repeated more times. Though our method was designed to use the PC algorithm, this does not exclude that the algorithm may change. In fact the user can substitute the PC algorithm with the algorithm that he prefers.

Novel aspects

The innovative contribution of this thesis is mainly related to the LGNs expansion task as mentioned above. In particular it is innovative how the task is treated (use of the LGN knowledge, type of gene expression data) and the type of obtained output.

- **Task and method**

The LGN expansion idea originates from two considerations. The first is that, often, a biological researcher has prior knowledge about relevant genes and their involvement in a LGN and he is willing to expand this knowledge. The expansion is obtained, at the beginning, with hypotheses formulation about other putative interactions of these genes with new genes and subsequently with the *in vivo* validation of this hypothesis by new experiments. The hypotheses can be formulated with bioinformatic systems (use of algorithms to infer regulatory networks, *in silico* coexpression analysis of gene expression data). The second consideration is that quite often the LGNs proposed by the commonly

used algorithms are complex, and thus it is difficult for a researcher to design the appropriate biological experiment to validate the results. In fact *in vivo* validation requires the characterization of all the genes of the new gene network. The techniques commonly used to measure gene expression on a large scale (such as microarray experiments) can not be used, because, often, they are the input of the algorithm. Other useful techniques to validate the *in silico* results are, for example, the chip-seq tecnique, which studies the binding sites of the transcription factor or the manipulation of a specific gene in the homologous or heterologous system (knock-out or over expression). In general these approaches can be used only to test few genes, because they require long times and they are labor intensive and costly.

In literature, a big number of articles try to infer new gene networks by identifying new causal relationships among genes [Penfold and Wild, 2011]. These articles describe algorithms, as were described in Chapter 2, or website platforms (also called web-based tools) which are big collections of different types of data (publications, information about gene annotation, gene expression and chemical data) used to reverse engineering putative gene-gene interactions. There are two principal classes of website platforms. The first class comprises the web-based tools that use prior knowledge about a GRN as scaffold and then use gene expression data to validate the relationships between genes (e.g. BAR [Toufighi et al., 2005], BioGRID [Stark et al., 2006], GeneMANIA [Mostafavi et al., 2008]). It is important to underline that the GRN scaffold derives from the published information and it may not contain all the gene-gene interactions because it may not represent the only true biological network that involves the studied genes. In fact, it is possible to miss gene-gene interactions that can be associated with specific phenotypes or specific development conditions, which are not present in any publication. It is also possible that the same genes are involved in more GRNs. The second class, instead, is represented by web platforms that use a combination of information deriving from public databases and gene expression data to infer GRNs (Predictive Networks [Haibe-Kains et al., 2012], bioPIXIE [Myers et al., 2005]).

The task of GRN expansion is recent and therefore few publications are available up to now. This fact highlights the importance and novelty of this topic. One of the first algorithms developed for GRN expansion was GENESYS [Tanay et al., 2001]. Subsequently other methods were proposed; namely Growing algorithm [Hashimoto et al., 2004], Gat-Viks and Ron Shamir [Gat-Viks and Shamir, 2007], BN+1 [Hodges et al., 2010], ANAP [Wang et al., 2012] and GENEIES [Kotera et al., 2012]. All these systems have in common with our method PC-IM the task, but differ in the approach used for the expansion.

GENESYS (GEnetic Network Expansion SYStem) [Tanay et al., 2001] is an algorithm that uses gene expression data to expand a known LGN. It proceeds with three

steps:

1. Standardization of the data of the input dataset;

2. Use of *a priori* information about LGN to obtain the fitness function. The fitness function is the criterion wich guides the selection of the expansion genes;

3. Expansion of the LGN analyzing a gene at a time and using the fitness function as selection criterion.

Growing algorithm [Hashimoto et al., 2004] uses the gene expression data to expand little LGNs (one or more genes) and prior knowledge about these LGNs is not mandatory. This method can be divided in two steps:

1. Measure of three parameters which reflect the strength of the connection between two genes. These parameters are measured between genes of the LGN, between the genes external to the LGN and between genes of the LGN and genes outside the LGN;

2. Combination of these three parameters in an unique criterion subsequently used to expand the LGN.

Gat-Viks and Ron Shamir [Gat-Viks and Shamir, 2007] in the 2007 have developed a system that uses gene expression data and prior knowledge to expand LGNs [Gat-Viks and Shamir, 2007]. Gene expression data must derive from experimental procedures (gene silencing or enhancement of the gene expression). There are three steps:

1. Modeling the prior knowledge. This implies that an evaluation model is created using *a priori* information about LGN, Bayesian scoring matrix and probabilistic modeling;

2. Generation of the predicted evaluation model from the experimental gene expression data;

3. Comparison of the predicted and observed evaluation model to discover the expansion genes of the LGN.

BN+1 [Hodges et al., 2010] is a system that uses the gene expression data and prior knowledge to expand a LGN. The steps are the following:

1. Generation of multiple cores Bayesian Networks (core BN) using gene expression data, prior knowledge of the LGN and the log posterior score;

2. Selection of the core BN with the highest log posterior score [Heckerman et al., 1995]. This core BN will be the LGN to expand;

3. Expansion of the core BN adding a gene at a time. The final expansion gene list will contain only those genes that improve the score determined in the second step.

GENIES [Kotera et al., 2012] discovers the new genes related with a specific LGN using a kernel function [Kotera et al., 2012]. It uses *a priori* information about LGN and different type of data in combination or alone (gene expression data, protein localization data, phylogenetic profile, kernel matrix based on the gene expression profile, kernel matrix based on the protein localization profile and kernel matrix based on the phylogenetic profile). It works in three steps:

1. Transformation of each data set in a kernel similarity matrix;

2. Mapping of the knowledge about LGN in a feature space (training process) equipeed with the Euclidean distance [Yamanishi et al., 2005].

ANAP [Wang et al., 2012] is a tool that was developed only for *Arabidopsis thaliana*. It integrates 11 Arabidopsis protein interaction databases, 100 interaction detection methods, 73 species that interact with Arabidopsis and 6.161 references [Wang et al., 2012]. This tool may expand the network only using the interaction detection methods present in its database, instead PC-IM and other methods listed above use the data loaded by the user (gene expression or other type of data).

- **Usage of the LGN prior knowledge**

Another innovative aspect of PC-IM is the step in which the prior knowledge of the LGN is used. All methods mentioned above use the prior knowledge at the beginning of the expansion process. Instead PC-IM uses *a priori* information in two different moments. At the beginning it uses, as prior knowledge, only the names (genes identifications) of LGN's genes to add them to the *tiles*. These genes will be the only genes present in all the subdivision of the input gene-list, instead the other genes will be present only in a single *tile*. Subsequently, the LGN knowledge will be used, at the end, to estimate the precision of the genes in the output expansion gene list. Practically, the genes of the LGN are treated as any other gene when applying PC-IM, while in the other expansion methods (GENESYS, Growing algorithm, Gat-Viks and Ron Shamir algorithm, BN+1, ANAP and GENIES), the prior knowledge of the LGN is used to construct a scoring matrix to be improved with the addition of the expansion genes.

- **Intrinsic performance evaluation**

PC-IM differs from the algorithms cited above also for the criterion used to select the final expansion gene list.

PC-IM uses the normalized frequency to determine which genes will be included in the expansion gene list of a LGN. The frequency corresponds to the number of times that a gene is found to expand a LGN with respect to the times that the same gene could be found. Each gene, not included in the LGN, can be present just once for iteration, namely it is present in only one *tile*. The frequency calculated on the LGN genes is used as a cut-off frequency to select the other genes.

GENESYS uses the fitness value. This is a numerical value that expresses the performance of an individual against other different individuals. In case of the expansion task the individuals that are presumed to have higher fitness values are the genes in the LGN, while the other individuals are the external genes [Liang et al., 1998]. **Growing algorithm** uses the strength of a connection to expand a LGN. This strength is determined from the coefficient of determination [Hashimoto et al., 2004].

"The coefficient of determination gives an indication of the degree to which a set of variables improves the prediction of a target variable relative to the best prediction in the absence of any conditioning observations "[Hashimoto et al., 2004].

Gat-Viks and Ron Shamir algorithm uses the Bayesian score [Gat-Viks and Shamir, 2007]. It is used as selection criterion of the model predictions to the data.

BN+1 uses the log of the BDe. This is the natural log posterior and its specific formulation is in Hodges et al. [2010].

GENIES uses the Euclidean distance to obtain the expansion genes. The Euclidean distance is calculated between genes of the LGN and between the hypothetical expansion gene and LGN genes. The Euclidean distance between the LGN genes is the threshold to select other genes [Kotera et al., 2012].

- **Gene expression data: observational data**

In PC-IM, the inference of the GRN is based on a particular type of gene expression data called observational data. This type of gene expression data is present in public databases, but it is rarely used for inference of networks. In fact there are two different strategies to determine GRNs from gene expression data. One relies on data measured in a perturbed biological system (experimental perturbed data) [Davidson et al., 2002], the other on the natural variation of expression levels of the same gene in different cells (observational data) [Chu et al., 2003] [Yoo et al., 2002].

The experimental approach is based on the suppression or the enhancement of the expression of one or more genes using transgenesis or natural mutants, and the measurement of how the gene expression is influenced [Davidson et al., 2002]. With this method

is relatively easy to identify the genes involved in the GRN and it has proved fruitful in unraveling small parts of a regulatory network. However, this approach has some disadvantages. It provides information only about the effects of one or few manipulated genes and its gene targets. Moreover, in complex organisms, such as human and plants, this type of data is difficult to obtain due to the long time to collect them and to ethical issues.

The second approach relies on observational data, and it overcomes these problems. In fact it permits to determine multiple relationships without any experimental intervention. The observational data are obtained merely observing a phenomenon (natural variation of the expression level) when the organism is in the natural (or optimal) development condition and when is under stress condition (water stress, cold stress, salt stress). In this case the GRN is inferred from statistical dependencies and independences among the measured expression level [Chu et al., 2003] [Yoo et al., 2002]. Despite the abundance of observational data present in the public databases only few algorithms use them [Spirtes et al., 2001] [Pearl, 2002] [Emmert-Streib et al., 2012], since they require elaborate statistical procedures.

The development of a system able to infer gene networks starting from observational data is of great importance since it allows to exploit the big availability of the data stored in public databases. This is an innovative aspect of this work. Moreover, this is also important because it allows to use public data and to find new genes involved in a specific LGN in a early stage of the biological investigation. This possibility offers the advantage of using new information to specifically design the experiment for novel genes validation.

- **Final output**

An innovative contribution of this thesis is the formulation of an alternative strategy to expand a Local Gene Network (LGN) by identifying only the additional genes that are related with at least one LGN-gene. In other terms PC-IM returns an expansion gene list, without specifying which genes of the LGN have a direct relationship with the genes in the expansion gene list.

In the expansion process of a LGN, as the inference of a GRN, there are two important steps. The first step regards the identification of the genes that expand the LGN. The second regards the identification of the causal relationships between the new genes and the LGN genes. Causal relationships are defined by the connection between two genes (relationship) and the orientation of this connection (causal direction). If y and x are two genes, and in particular y is affected by x (the presumed cause), there are three conditions that determine the exact causality (direction of the relationship) between these two genes [Kenny, 1979]:

1. x must precede y temporally;

2. x must be reliably correlated with y (beyond chance);

3. the relation between x and y must not be explained by other causes.

Nevertheless these three conditions are necessary, but not sufficient to find all possible causal relationships. This because there are cases (reciprocal causation and simultaneous causation) that are not explained by these conditions [Antonakis et al., 2010].

Reciprocal causation: in gene networks, there is a possibility that two variables are reciprocally cause and effect. This occurs when the expression of gene x activates gene y and the encoded protein inhibits the expression of gene x. A biological example is the lac operon [Van Hoek and Hogeweg, 2007]. In the lac operon an increased concentration of lactose in the cell causes an increase of the lac operon activity. Simultaneously an increase of the activity of this operon decreases the lactose cellular quantity.

Simultaneous causation is the activation of a target gene by the action of more genes together. For example this happens in tryptophan synthesis. In this case a trypthophan operon controls the synthesis of the enzymes that produce tryptophan. The operon is regulated by a repressor that alone is inactive, but is induced when it combines with a specific molecule [Hamon et al., 1981].

These two considerations show that it is very difficult to find by *in silico* analysis the exact causal relationship between two genes. It is easier to identify the list of genes that expands a LGN. Moreover this information (the list of genes) is enough to design *in vivo* validation experiments. The experiments will help in validating the new genes and to discover the causal relationships between the genes and the LGN genes. For the above reasons we choose the expansion task, (focusing on the list of the expansion genes and not on the causal relationship), rather than discovery of new GRNs.

1.2 Structure of the Thesis

The thesis is composed of four main Chapters. Chapter 2 presents an overview of the different types of reverse engineering algorithms in Section 2.1 and a detailed description of the PC algorithm and its related applications in Section 2.2. Section 2.3 a description of the methods used to expand LGN is reported. The other three Chapters describe the results of my PhD work. Chapter 3 reports the description of PC-Iterative Method (PC-IM). Chapter 4 presents the Evaluation of PC-IM. It starts with a section dedicated to a preliminary evaluation in which the expression data for testing the performance of the method are selected (Section 4.1) and the choice of the PC algorithm is motivated (Section 4.2). Afterwards in Section 4.3 an evaluation of the PC-IM is presented. This includes the assessment in terms of the performance of the single parts of our method.

Finally Chapter 5 is focused on the real expansion of a specific Local Gene Network with PC-IM and two case studies are described.

Chapter 2

State of the art

Causation is a relationship between an event (the cause) and another event (the effect). The second event (the effect) is interpreted as a consequence of the first event and it can have more than one cause [Spirtes et al., 2001]. *Causation* has three properties; it is transitive, irreflexive and antisymmetric.

- if X is a cause of Y and Y is a cause of Z, then X is also a cause of Z (transitive property);
- an event X cannot cause itself (irreflexive property);
- if X is a cause of Y then Y is not a cause of X (antisymmetric property).

Causal inference is the process used to obtain conclusions about presence/absence of causal relationships between events. To draw to these conclusions statistic means are used [Spirtes et al., 2001]. To represent causality we can use a directed graph. A directed graph consists of a set of vertices (e.g. genes) and a set of directed edges (e.g. relationship between genes), where each edge is an ordered pair of vertices. In Figure 2.1 an example of a directed graph G1 is depicted and below we report the terminology associated to the graph.

- the *vertices* are {A, B, C, D, E};

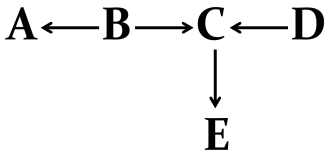


Figure 2.1: Example of a directed graph G1.

- the *edges* are {B → A, B → C, D → C, C → E};
- B is *parent* of A, because there is an oriented edge from B to A;
- A is *child* of B;
- A and B are *adjacent*, because there is an edge between the two variables;
- a *path* is a sequence of adjacent edges (A ← B → C);
- a *directed path* is a sequence of adjacent edges all pointing in the same direction (B → C → E);
- C is *collider* on the path because both edges on the path are directed into C;
- E is *descendent* of B (and B is an *ancestor* of E), because there is a directed path from B and E. Each node is *ancestor* and *descendent* of itself.

When a directed graph does not have cycles, then it is called directed acyclic graph (DAG). This means that in the DAG there is no directed path from any vertex to itself.

The causality representation by directed graph and/or DAG presents some problems. For example in the graph G1 considering A → C ← B we are not able to represent the situation in which there are two different drugs (A and B) that reduce symptom C, and A can reduce the symptom C also without B, instead B alone has no effect on C. Moreover we are not able to represent the situation in which there are two independent variables A and B with two states. For example A is a battery and the two states are charged and uncharged; B is a switch and two states are on, off. A and B cause C (A → C ← B), only when A and B are simultaneously verified (A indicates the battery charged and B the switch on) [Spirtes et al., 2001]. These problems arise, because the relationships are represented through the probability distribution associated with the graph [Spirtes et al., 2001].

As mentioned in Chapter 1, the use of the transcript levels to identify regulatory influences between genes is called reverse engineering (or inverse modeling or network inference). There are two classes of reverse-engineering algorithms: those that search for physical interactions and those which search for influence interactions [Gardner and Faith, 2005]. The aim of the physical interaction's methods is to identify the binding motifs of transcription factors and identify thus their target genes (gene-to-sequence interaction). Influence interaction methods, instead, seek to relate the expression of a gene to the expression of the other genes in the cell (gene-to-gene interaction). In this work the ensemble of these influence interactions constitute a gene network. In this section we present a brief review of the principal algorithms developed to find influence interactions.

Obtaining a gene network from influence interactions is useful for multiple purposes:

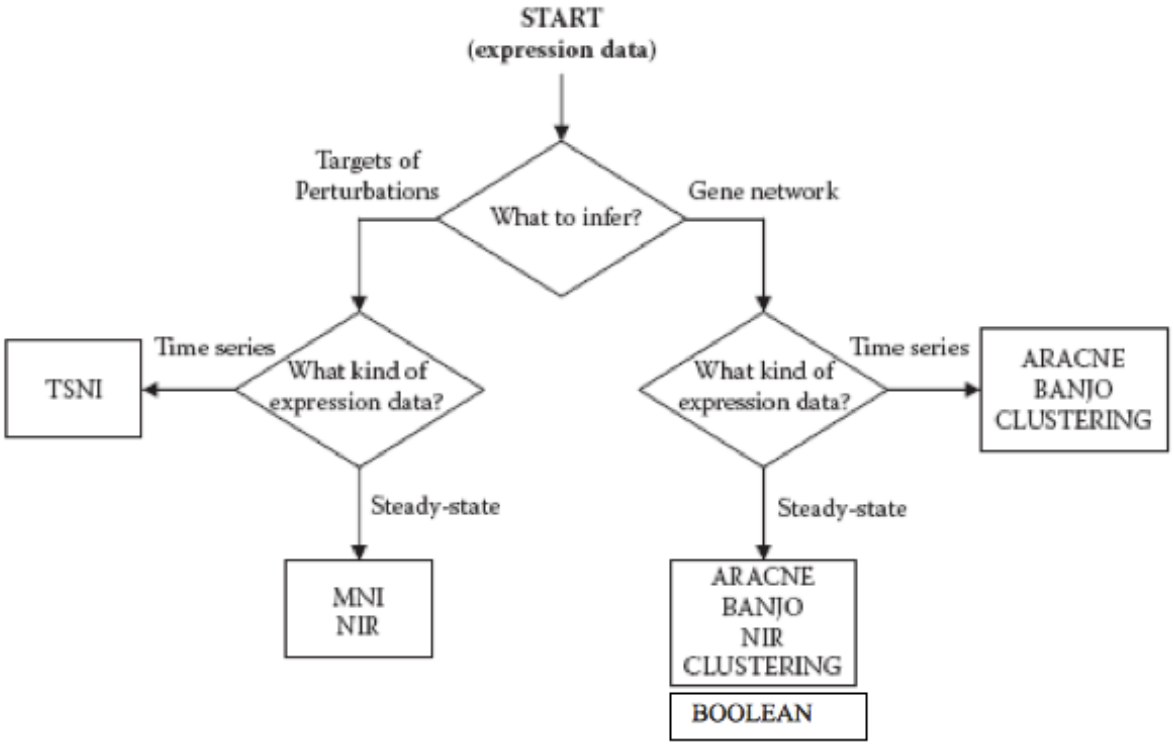


Figure 2.2: **Classification of different algorithms based on their specific domain of application** (adapted from Lauria and di Bernardo [2010])

1. identification of the genes that regulate each other with multiple (direct and/or indirect) interactions;

2. prediction of the response of a network to perturbations;

3. identification of the real physical interactions. This identification is obtained by the integration of the gene network with additional information from sequence data and other experimental data (i.e. chromatin immuno-precipitation [Das et al., 2004] or yeast two-hybrid assay [Bartel and Fields, 1997])

2.1 Gene network inference algorithms: a review

PC-IM was developed with the PC algorithm, but theoretically it can be used also with other algorithms. For this reason in this section a review of the main algorithms used to infer gene networks is included.

Figure 2.2 shows the different domains of application of the most used algorithms according to the type of experiments that have generated the input data.

In the algorithm's description we will use the following variables:

i are the genes;

N is the total number of genes;

x_i is the expression measurement of gene i;

X is the set of expression measurements for all the genes;

M is the total number of time points (or different conditions) of the expression measurements;

a_{ij} is the interaction between gene i and j.

In the undirected graph, the direction is not specified and $a_{ij} = a_{ji}$, instead when $a_{ij} \neq a_{ji}$ we have a directed graph. A directed graph can also be labeled with a sign and strength for each interaction $a_{ij} > 0$ means that there is activation, instead $a_{ij} = 0$ means there is no interaction and $a_{ij} < 0$ indicates the repression). The choice of the type of graph (directed or undirected) depends on the inference algorithm.

2.1.1 Clustering algorithms

Clustering algorithms divide the genes in groups (clusters). In each group there are genes with similar expression profiles (coexpressed genes). The coexpression between genes does not imply, however, the direct interaction among these genes [Lee et al., 2004]. In fact genes that are coexpressed can be related together by one or more intermediaries (indirect relationships). For this reason the clustering algorithms are not properly network inference algorithms, but are rather used to visualise and analyse gene expression data. Moreover the coexpression analysis can be used to deduce the function of genes from other genes in the same cluster [Eisen et al., 1998].

The most common clustering algorithm is the hierarchical clustering [Eisen et al., 1998]. It searches to obtain a single tree where the branch lengths reflect the degree of similarity between the genes. The connection between genes is assessed by a pairwise similarity function (for example Pearson correlation). The highest value of the pairwise correlation coefficient indicates that there is a relationship between the pair of genes [Eisen et al., 1998].

Another algorithm is the signature algorithm [Ihmels et al., 2002]. It is specialized to identify transcription modules starting from gene expression data. The transcription module is a group of genes that are co-regulated in particular experimental conditions. This algorithm enables gene classification, namely a clusterization of genes into different groups [Ihmels et al., 2002].

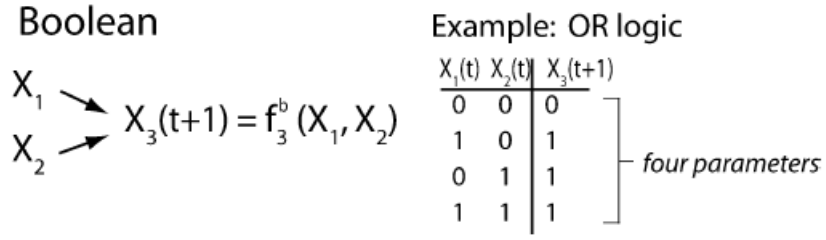


Figure 2.3: **Boolean model used to represent the relationship between input and output transcripts** (taken from Gardner and Faith [2005]).

2.1.2 Network Inference Algorithms

The inference of GRNs from expression data is a difficult task, mainly because the number of variables is much larger than the number of observations. Following a description of the most common algorithms developed to this aim.

Boolean network models

Kauffman proposed the first Boolean model [Kauffman, 1969]. They represent genetic networks as interconnected binary elements, with each of them connected to a series of others [Kauffman, 1969]. In the GRN the binary elements are the genes and each element can have two binary states, inactive (0) or active (1) and the interaction between the elements are modeled as Boolean rules. This mean that, if fully connected, a Boolean network with N genes will present 2^N gene expression patterns. This number is very high and requires large amount of experimental data [Gardner and Faith, 2005]. For this reason, it is assumed that networks are sparsely connected and this shows the importance to specify the connectivity. Moreover the Boolean network can be represented as a directed graph, where the edges are represented by Boolean functions (simple Boolean operations, e.g. AND, OR, NOT). In general a repressor is equivalent to a NOT function, whereas cooperatively acting activators are represented with the AND function. In this way, the Boolean variables (where the states of genes can be 1 or 0) are determined at time t+1 by the state of the network at time t, the value of the K inputs and the logical function assigned to each gene. Figure 2.3 is an example of Boolean model in which the OR logic function is illustrated. In this representation there are three transcripts X_1, X_2 and X_3 and K has value 2. This implies that the possible networks, with the three variables, are 4.

The final aim of the Boolean method is to identify a Boolean function for each gene in the network that it explains the model. This class of algorithms have been used to describe different biological pathway, including signalling pathway [Shymko et al., 1997]

[Genoud et al., 2001] and bacterial degradation processes [Serra and Villani, 1997]. In the literature various Boolean algorithms have been proposed. One of this is REVEAL (REVerse Engineering ALgorithm; [Liang et al., 1998]). "REVEAL was developed to allow for multiple discrete states as well as to let the current state depend not only on the prior state but also on a window of previous states"[Hecker et al., 2009].

Bayesian network models

Bayesian Network (BN) models are graphical representations of joint probabilistic distributions of a set of random variables X_i (i.e. gene, protein or other cellular elements). A BN has two components [Pe'er, 2005], the first component is a DAG that represents the relationships between the variables. Its vertices are the random variables X_i and the edges represent the influence of one variable on another. The second component, denoted with θ describes a conditional probability distribution for each variable X_i.

The principal limitation of the BN is the absence of cycles in the network as well as the explicit treatment of causality among the variables. The absence of cycles derive from the use of the DAG to represent the network. This is a problem, because the cycle systems are relevant in biological systems (for example the feedback loops among the B-C variables in Figure 1.1). The other issue is that BNs represent the probabilistic dependencies among variables and not causality. This mean that the parents of a node are not necessarily also the direct causes of its behaviour, in our case its gene expression [Bansal et al., 2006]. To overcome these limitations the Dynamic Bayesian Network (DBN) was developed [Yu et al., 2004]. DBNs can establish the direction of causality because they incorporate temporal information [Yu et al., 2004], but they need a large quantity of input data, such as gene expression data. These data, in molecular biology, is often limited, in particular for complex organisms.

An algorithm based on the Bayesian Network formalism is Banjo (Bayesian Network Inference with Java Objects). It implements both BN and DBN [Yu et al., 2004]. Hartemink and colleagues developed Banjo [`http://www.cs.duke.edu/~amink/software/banjo/2008-06-20`] [Yu et al., 2004]. The output of Banjo is a signed directed graph indicating regulation among genes. To this aim Banjo infers the parameters of the conditional probability density distribution for each network structure explored. An overall network's score is computed using the scoring metric Bayesian Dirchlet equivalence (BDe) in the Banjo's Evaluator module. At the end the output network will be the one with the best score (Banjo's Decider module) [Bansal et al., 2006].

Differential Equation Model

The Differential Equation Model is a deterministic approach that describes gene regulation as a function of other genes in terms of Ordinary Differential Equations (ODEs). To reach this aim a set of ODEs are provided for each gene. In the set of ODEs, each equation

describes the variation in time of the concentration of a particular transcript, x_i, as a non linear function f_i of the concentrations of the other transcripts [Gregoretti et al., 2010]:

$$\begin{aligned} x_i(t) &= f_i(\underline{x}_i(t), u; \theta) \\ \underline{x}_i &= [x_i, ..., x_N] \\ i &= 1...N \end{aligned} \tag{2.1}$$

Where t is the time in which the transcripts are measured, $\underline{x}$(t) is a vector whose components are the concentrations of the transcripts x_i measured at time t, u_i(t) is the external perturbation applied at gene i at time t, θ is a set of parameters describing the interactions between genes. With this system the edges, among the variables, represent causal interaction, and not statistical dependencies as the other methods [Bansal et al., 2006].

This type of reverse-engineering algorithms is used to reconstruct gene-gene interaction starting from the steady state of gene transcript concentration (i.e. RNA expression measurements or time series measurements) and its subsequent external perturbation.

Two algorithms based on ODE are Network Identification by multiple Regression (NIR [Gardner et al., 2003]) and Microarray Network Identification (MNI [di Bernardo et al., 2005]). Both are based on the same equation [Bansal et al., 2006]:

$$\sum_{i=1}^{N} a_{ij} x_j = -b_i u \tag{2.2}$$

if b_i represents the effect of the external perturbation on x_i and there are M time points, then this equation 2.3 derives from the equation:

$$x_i t_k = \sum_{j=1}^{N} a_{ij} x_j(t_k) + b_i u(t_k) with \begin{cases} k & = 1...M \end{cases} \tag{2.3}$$

when the case of steady-state data and $x_i(t_k) = 0$ the i-th gene becomes time independent.

The NIR supposes that the data $\underline{x}$ (transcript concentrations) and u (the perturbation) are normally distributed with known variance [Gregoretti et al., 2010]. It uses, as input data, the gene expression data following each perturbation experiment and the knowledge of which genes have been directly perturbed in each perturbation experiment [Gardner et al., 2003].

MNI algorithm needs, as input data, microarray experiments that are a result from any kind of perturbations. MNI does not require knowledge of $b_i u$ [Bansal et al., 2006].

The particularity of MNI is that it uses the inferred network to filter the gene expression profile after a treatment with a compound, to determine pathway and genes direct target of the compound. For the details of NIR and MNI algorithms see di Bernardo et al. [2005].

Another ODE algorithm is the Time Series Network Identification (TSNI [Bansal et al., 2006]). TSNI identifies the gene network when the gene expression data are dynamic. This mean that, unlike NIR and MNI, $x_i(t_k) \neq 0$ and M time points following the perturbation are measured. The complete description of the TSNI is presented in Bansal et al. [2006].

Information theoretic approach (Association networks)

Information theoretic approaches assign interactions to pairs of transcripts that exhibit high statistical dependence in their responses in all experiments in a training data set. To measure dependence, the two most common strategies are Pearson correlation and Mutual Information (MI). The Pearson correlation assumes linear dependence between variables, instead MI measures the degree of dependence between two variables (genes). In fact given two variables, MI determines the ratio between the probability to find two variables together with the probability to find each variable individually [Fernandes and Gloor, 2010]. Mutual Information MI_{ij} between genes i and j is computed as:

$$MI_{ij} = H_i + H_j - H_{ij} \tag{2.4}$$

where H is the entropy and it is defined as:

$$H_i = -sum_{k=1}^{n} p(x_k) \log(p(x_k)) \tag{2.5}$$

the higher the entropy the more the gene expression levels across the experiments are randomly distributed. To find the LGN, MI is computed for each pair of genes and its value is included into the range [0,1]. Higher value of MI (value close to 1) indicate that two gene are non-randomly associated to each other [Bansal et al., 2006]. The value of MI becomes 0 when two variables x_i and x_j are statistically independent. MI is more general than Pearson correlation coefficient but this property does not prevent to get almost identical results [Steuer et al., 2002]. In the information theoretic approaches, the edges in the network represent only a statistical dependency and not a direct causal interaction between the variables.

ARACNE

The Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) was developed for the reverse engineering of human trascriptional networks from gene expression data [Basso et al., 2005] [Margolin et al., 2006]. In particular is was developed for the reconstruction of trascriptional networks of human B cells [Basso et al., 2005] [Margolin et al., 2006]. Subsequently ARACNE has been also used to predict metabolic network from high throughput metabolite profiling data [Nemenman et al., 2007]. ARACNE assumes that each gene expression level is a random variable and the mutual relationships

between pairs of variables can be obtained by statistical dependencies. In this way it defines an edge as an irreducible statistical dependency between gene expression profiles.

This algorithm can be divided in two main steps at the and the output is an adjacency matrix, namely a matrix which reports the candidate interactions.

Step 1: identification of candidate interactions by estimating Mutual Information (MI) (Equation 2.6) for all pairs of gene in the geneset, $I(g_i, g_j) = I_{ij}$. This is an information theoretic measure of relatedness that is zero if the joint distribution between the expression level of gene i and j satisfies $P(g_i, g_j) = P(g_i)P(g_j)$. Then ARACNE excludes all the pairs for which the null hypothesis of mutually independent genes cannot be ruled out.

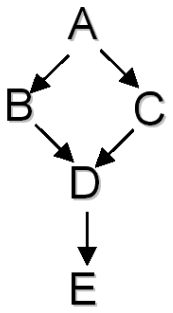
Step 2: ARACNE filters the statistical dependencies, eliminating those with MI values below the appropriate threshold I_0. This allows for removing the most indirect candidate interactions using a know information theoretic approach: the Data Processing Inequality (DPI). DPI is a property of MI that states if gene g_1 and g_3 interact only by another gene (g_2), then $I(g_1, g_3) \leq min(I(g_1, g_2); I(g_2, g_3))$ [Cover and Thomas, 2006]. This implies the removal of the least one of the three MIs, because it can come only from indirected interactions.

2.2 The PC algorithm

The PC algorithm tries to find the causal relationships between the variables. Peter Spirtes and Clark Glymour developed the algorithm for the social science domain and its name comes from their names (PC: Peter and Clark) [Spirtes and Glymour, 1991]. It assumes a Bayesian causal network model and it makes use of valid statistical testing to produce a DAG as output. It comprises three steps. In the first step it applies the conditional independent test to discover relationships between variables. In the other steps it tries to orientate these relationships without creating cyclic structures. Before describing in details the PC algorithm and considering its modifications it is necessary to introduce some preliminary definitions.

Causally sufficient criteria

A set of variables $\mathbb{V}$ is *causally sufficient* when no two members of $\mathbb{V}$ are caused by a third variable both in $\mathbb{V}$. Zhang and Spirtes, [Zhang and Spirtes, 2008], emphasize that "the idea, of the causally sufficient criteria, is that X is direct cause of Y relative to the given set of variables when it is possible to find some pair of interventions of the variables other than Y that differ only in the value they assign to X but will result in different post-intervention probability of Y"[Zhang and Spirtes, 2008]. With the sentence "X is cause of Y"we mean that an intervention on X, makes a difference to the probability of



A.

B.

Variable	Parents	Descendants
A	Ø	A, B, C, D, E
B	A	B, D, E
C	A	C, D, E
D	B, C	D, E
E	D	E

Figure 2.4: **Representation and classification of the variables of a DAG G.**
A. representation of the variables of a DAG G. **B.** classification of the variables of a DAG G.

Y.

For example Figure 2.4-A shows a representation of a DAG G and Figure 2.4-B a table with the classification of the variables of G. The DAG G has 5 variables (A, B, C, D and E) and the arrowheads between the variables are oriented edges. In G (Figure 2.4-A) the set B, C, D, E is not *causally sufficient*, because there is another variable A, not included in the set, which is a direct cause of the B and D variables.

Conditional independence

The definition of conditional independence is as follows: "Two random variables X and Y are conditional independent given a set $\mathbb{Z}$ of variables on distribution P, written as $I_P(X, Y|\mathbb{Z})$, if $P(X|Y, \mathbb{Z}) = P(X|\mathbb{Z})$ and $P(X|\mathbb{Z}) \neq 0$, $P(Y|\mathbb{Z}) > 0$, where $P(X|\mathbb{Z})$ means the conditional probability of X given $\mathbb{Z}$. In an other way we can say that X and Y are conditional, by independent when $P(X|Y, \mathbb{Z}) = P(X|\mathbb{Z})P(Y|\mathbb{Z})$. This mean that if X and Y are independent conditioned on the $\mathbb{Z}$, then does not provide any information about Y once given knowledge of $\mathbb{Z}$ and *vice versa* [Spirtes et al., 2001]."

With only the causally sufficient criteria it is very difficult to find the best DAG from a given sample, since the number of possible DAGs is greater than the exponential of the number of observed variables. To reduce the number of possible DAGs, the Bayesian models use together other two different assumptions: Causal Markov Condition (CMC) and Causal Faithfulness Condition (CFC).

Causal Markov Condition (CMC)

The CMC states that given a set of variables whose DAG G represents the causal structure of these variables, each variable is independent of its non-descendents conditional on its directed causes (its parents in graph G) [Ramsey et al., 2012]. In particular in Figure 2.4-A, the CMC entails that if there is no edge between two variables A and D in a DAG G, then A and D are conditional independent on some subset of the other variables $\mathbb{Z}$

($\mathbb{Z} = B, C$) ($I_P(D, A|B, C)$. In addition to the example between the variables A and D there are other conditional independence relations entailed by CMC:

$I_P(A, \emptyset|\emptyset)$;

$I_P(B, C|A)$;

$I_P(C, B|A)$;

$I_P(E, \{B, C, A\}|D)$

These relations may originate other conditional independence relations; for example $I_P(E, A|\{B, C\})$. Another interesting consideration is that the CMC alone implies the principle of the common cause. In fact if two variables X and Y are not conditionally independent on $\emptyset(\sim (X, Y|\emptyset))$, then, according to the CMC, we have three possibilities: X is cause of Y, or Y is cause of X, or it exists a third variable that is the common cause of both X and Y (common cause).

Causal Faithfulness Condition (CFC or Stability Condition)
The CFC says that given a set of variables $\mathbb{V}$ and DAG G is the causal graph, DAG G is the true causal graph when it is the exact map of the distribution probability ($P_{\mathbb{V}}$) of the variables in the set $\mathbb{V}$. The probability distribution P entailed by a causal graph G satisfies the CFC if and only if every conditional independence relation true in P is entailed by the CMC applied to G [Zhang and Spirtes, 2008]. Under the CFC, conditional independence relations give direct information about the structure of the graph. In Figure 2.4-A with the CFC we can conclude that there is no direct edge between A and D if a statistical test indicates that A is independent of D conditional on $\mathbb{Z}(\mathbb{Z} = \{B, C\})$ [Zhang and Spirtes, 2008].

Assuming together the CMC and CFC it is possible to reduce the total number of DAGs, because they entail that conditional independency holds in the population if and only if the true causal DAG entails it by application of the Markov condition. To explain this concept we suggest to see the example present in the paper by Zhang and Spirtes [2008] (Figure 1 from paper Zhang and Spirtes [2008]). Given the DAG G in Figure 2.4-A, the CFC entails that $I_P(D, A|\emptyset)$.

d-separation
There is a method to ascertain whether the CMC and/or CFC entail conditional independence relation. The method is called *d-separation* (*d* means dependence), it is a graph-theoretical approach and is defined as follows.

Two variables X and Y are *d-separated* by a node set $\mathbb{Z}$ if and only if every path between X and Y is blocked. A path is blocked when there is an intermediate variable $Z \in \mathbb{Z}$ such that:

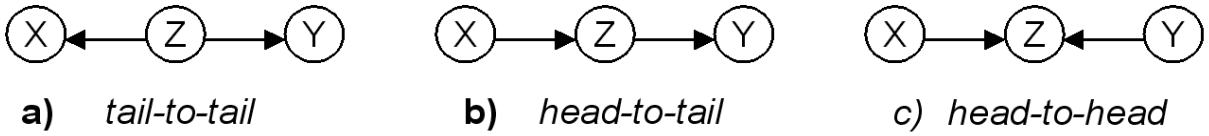


Figure 2.5: **Different types of connections considered in the *d-separation* step.**

1. the connection through Z is *tail-to-tail* or *head-to-tail* and Z has received evidence, or;

2. the connection through Z is *head-to-head* (or v-structure) and neither Z or any of Z's descendants have received evidence.

The different types of connections are represented in Figure 2.5 in which each nodes of the different graph represent a variable and each arrowhead is an oriented edge.

2.2.1 Description of the PC algorithm

The PC algorithm reconstructs the causal structure of the variables described by the input data, starting from the assumptions of the Causal Markov Condition, faithfulness and causal sufficiency of a graph. The PC algorithm works by progressively removing the edges from a complete undirected graph built on the variables given in the input data, until no more edges can be deleted, according to a function that decides when to delete the edge. The graph so obtained is called skeleton and it is then oriented according to the *d-separation* rules.

The PC algorithm receives a set $\mathbb{V}$ of random variables in input and it works in three phases described in the pseudo-code in Figure 2.6 and in the representation in Figure 2.7.

Phase 1: find the skeleton by deleting edges between independent variables
The PC algorithm starts generating a complete undirected graph G' from the set of variables V. Each node in the G' is a variable of V thus from now the variables will be called also nodes. Subsequently, the PC algorithm starts to remove the edges in G' testing the set of Adj(X). The idea is that if the set of independences is faithful to a graph, then there is not a link between variables X and Y, if and only if there is a subset S of adjacent nodes of X (Adj(X)) such that $I(X, Y|S)$ [Spirtes et al., 2001]. For each pair of variables in the subset S, $S_X - Y$ will contain such a set, if it is found.

In particular in this Phase the PC algorithm uses the Partial Correlation Coefficient (PCC) to estimate conditional independencies. This parameter corresponds to the correlation coefficient between the dependent and independent variables when all the effect of the other variables are removed [Kalisch and Bühlmann, 2007].

PC Algorithm

Phase 1: find the skeleton by deleting edges between independent variables

Form a complete undirected graph G' from the set of nodes $\mathbf{V}$
$n = 0$
repeat
  for all $X \in \mathbf{V}$ **do**
    for all $Y \in \mathbf{Adj}(X)$ **do**
      repeat
        choose an $\mathbf{S} \subseteq \mathbf{Adj}(X) \backslash Y$ and $\mid \mathbf{S} \mid = n$
        if $I(X, Y \mid \mathbf{S})$ **then**
          delete $X - Y$ edge from G'
          save $\mathbf{S}$ in $\mathbf{SepSet}(X, Y)$ and $\mathbf{SepSet}(Y, X)$
        end if
      until $X - Y$ edge is deleted or all $\mathbf{S}$ have been chosen
    end for
  end for
  $n = n + 1$
until $\mid \mathbf{Adj}(X) \backslash Y \mid < n$

Phase 2: orient all possible v-structures

for all triple $X - Z - Y$ where $X \notin \mathbf{Adj}(Y)$ **do**
  if $Z \notin \mathbf{SepSet}(X, Y)$ **then**
    orient triple as $X \rightarrow Z \leftarrow Y$
  end if
end for

Phase 3: orient remaining unoriented edges using rules

repeat
  for all $X \rightarrow Y - Z$ where $X \notin \mathbf{Adj}(Z)$ **do**
    orient $X \rightarrow Y \rightarrow Z$
  end for
  for all $X - Y$ such that there is a *directed* path from X to Y **do**
    orient $X \rightarrow Y$
  end for
until no more edges can be oriented

Figure 2.4: pseudocode of the PC-Algorithm.

Figure 2.6: **Pseudocode of the PC algorithm.**

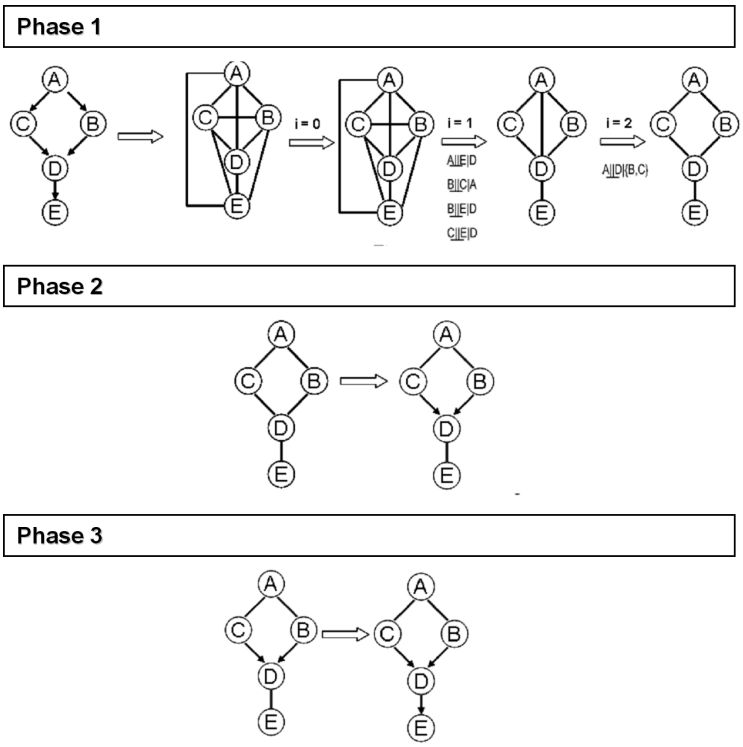


Figure 2.7: **PC algorithm schematic representation.**

Phase 2: orient v-structures (*head-to-head*)
The orientation of the edges in G' proceeds by examining sets of three variables $\{X, Y, Z\}$ such that in G' there are the unoriented links between X and Z and between Y and Z, but the link between X and Y does not exists. Then if Z is not included in SepSet(X,Y), the PC algorithm orients the edges from X to Z and from Y to Z creating a *v-structure* (*head-to-head*): $X \rightarrow Z \leftarrow Y$ [Spirtes et al., 2001].

Phase 3: orient the remaining unoriented edges using rules
In the Phase 2 not all link between nodes are oriented, so in this phase the PC algorithm tries to orient the rest of the edges. To arrive to this aim it follows two rules:

- Cycles have to be avoided;

- New v-strucuters have to be avoided.

2.2.2 Proposed modifications of the PC algorithm

The PC algorithm as such was applied on gene expression data [Wimburly et al., 2003] and more recently it has been improved in its different Phases (Figure 2.8). In Phase 1 PPC was substituted with Conditional Mutual Information by Zhang et al. [2012] and

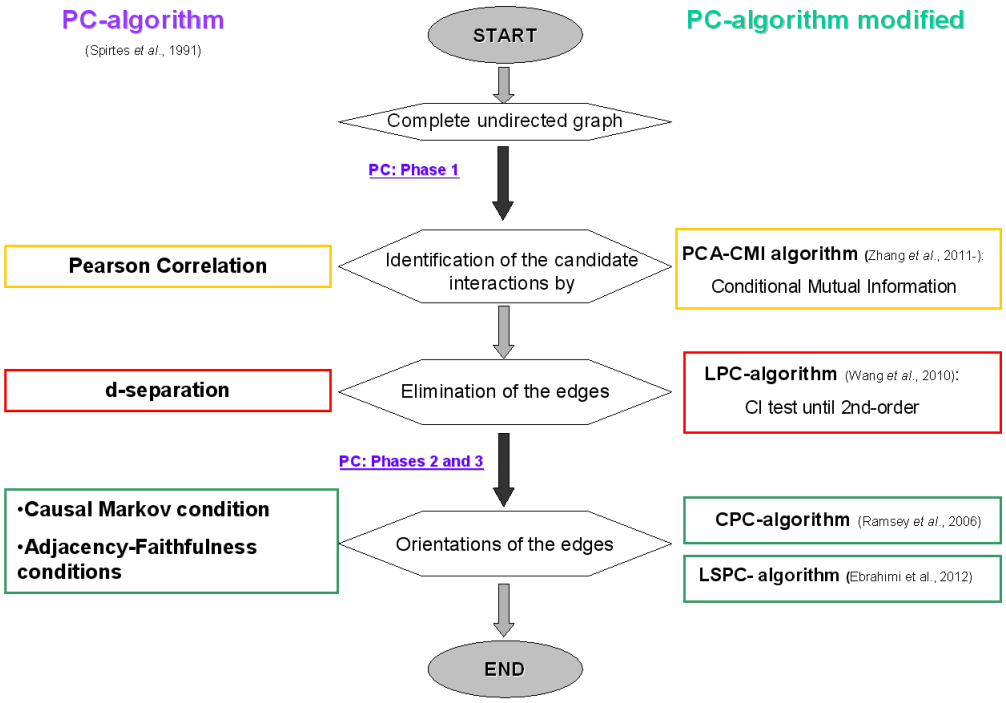


Figure 2.8: **Schematic representation of the differences between the original PC algorithm and its modified versions.**

the way in which the interactions between the nodes in the complete undirected graph are removed has changed [Wang et al., 2010]. In Phases 2 and 3 the way in which the edges are oriented was changed [Ramsey et al., 2012] [Ebrahimi et al., 2012]. In this section all these versions of the PC algorithm are reviewed.

Conservative PC algorithm (CPC) [Ramsey et al., 2012]

This algorithm aims to improve the PC algorithm in the orientation phase [Ramsey et al., 2012]. Ramsey highlights how the CFC assumption is formed from two components Adjacency-Faithfulness and Orientation-Faithfulness and how the causal Markov and Adjacency-Faithfulness conditions fail to orient the edges between the variables. An example shows this fact: consider three variables $< A, B, C >$ where A is independent from C (A$\perp$ C) and A$\perp$C|B (A $\rightarrow$ B $\rightarrow$ C). In this situation the Causal Markov and Adjacency-Faithfulness are both satisfied, but Orientation-Faithfulness is not true for this triple.

The PC algorithm removes the edge between A and C, because $A \perp C$, but orients the edges in this way A$\rightarrow$B$\leftarrow$C, because B is not in SepSet found in Phase 1. To overcome this problem CPC algorithm in Phase 2 tests for each triple $< A, B, C >$ which are the potential parents of A and C and not which are collider or non-collider. The Phase 2 in CPC is as follows. Let G a graph resulting from Phase 1, for each unshielded triple

$< A, B, C >$, check all subsets of A's and C's potential parents:

a. if B is not in any set conditioned on which A, and C are independent, orient A-B-C as A→B←C;

b. if B is in all sets conditioned on which, A and C are independent, leave A-B-C as it is;

c. otherwise, mark the triple as "unfaithful"by underlining the triple. This mean that there are possible different DAGs (A→B→C, A←B→C, A←B←C, A→B←C).

Low PC algorithm (LPC) [Wang et al., 2010]

This algorithm was developed to make easier the application of the PC algorithm on large gene expression datasets. In fact the PC algorithm requires an high number of tests, because all possible combinations of the conditioning set have to be examined. For this reason LPC uses the procedures of the PC algorithm, but it executes only the low-order Conditional Independence (CI) tests. In fact, in LCP, the number of CI tests is limited by the k specified by the user. The limited order of the CI tests reduces the computational complexity, but does not improve the sample size to analyse. In fact both these two algorithms (PC algorithm and LPC) have the best performances with sample sizes of 100 and 1000 variables [Wang et al., 2010].

LCP has two phases: CI test and orientation phase. In the first phase a limited number of CI tests is executed in comparison to the PC algorithm. The number of CI tests depends on k and the value of k is given as input data together with the dataset D (e.g. microarray with n genes an m measurements). In the second phase (orientation phase), the neighbor number of connected node pairs is checked before applying orientation rules, because the neighbor number of connected nodes is linked with the k value. If the k value is equivalent to n-2, where n is the number of genes in the input dataset D, the LPC algorithm is equivalent to the PC algorithm. Therefore we can say that LPC is a generalization of the original algorithm (in the sense that it constrains the search with ao additional parameter) and not a variation.

Path Consistency Algorithm with Conditional Mutual Information (PCA-CMI) [Zhang et al., 2012]

PCA-CMI is a method used to infer GRNs from gene expression data based on the PC algorithm, but it substitutes PCC [Kalisch and Bühlmann, 2007] with Conditional Mutual Information (CMI) [Zhang et al., 2012].

Many GRN inference algorithms are based on Mutual Information (MI). They start by computing the pairwise MI between pairs of genes, then the MI values are elaborated to identify the regulatory relationships [Altay and Emmert-Streib, 2010] [Fernandes and

Gloor, 2010]. In particular for two discrete variables X and Y, MI measures the dependency between X and Y and is defined as:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.6}$$

In case the two variables are independent $p(x,y) = p(x)p(y)$.

MI presents the advantage of measuring non-linear dependency (more common in biology) and it is able to deal with thousands variables in the presence of a limited number of samples [Meyer et al., 2007]. The problem is that MI is able to test pairs of genes not considering that there are more than two co-regulators. To overcome this problem Zhang et al. [2012] have proposed Conditional Mutual Information (CMI). This parameter is able to identify the joint regulations by exploiting the conditional dependency between genes of interest. CMI, in fact, is the expected value of the Mutual Information between two variables X and Y, given that a third variable Z or a set of variables Z has occurred. It can be defined as:

$$CMI(X,Y|Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \tag{2.7}$$

where $p(x,y,z)$ indicates the joint probability.

Limited Separator set in the PC algorithm (LSPC) [Ebrahimi et al., 2012] The LSPC algorithm aims to improve the way in which the edges are oriented in the Phase 2 and 3 of the PC algorithm. The main difference between these two algorithms is the choice of the separator set between the nodes of the graph G resulting from the Phase 1 of the PC algorithm.

PC considers as separator set of two vertices X and Y, all nodes that are present in the Adj(X) and Adj(Y). For LSPC the separator set is formed from all variables mostly repeated in the walks between X and Y. This method appears to improve the PC algorithm, because it reduces statistical errors in the step of edges orientation [Ebrahimi et al., 2012].

2.3 Methods for network expansion

GENESYS (GEnetic Network Expansion SYStem) [Tanay et al., 2001] is an algorithm that computes the fitness function of the LGN and then adds genes and relationships to find an expansion of the LGN that improves the fitness (Figure 2.9). In this system a biological network (or model) is defined from a set $\mathbb{U}$ of variables (e.g. genes or proteins), a set $\mathbb{C}$ of values (states) that the variables may attain, and functional dependence between

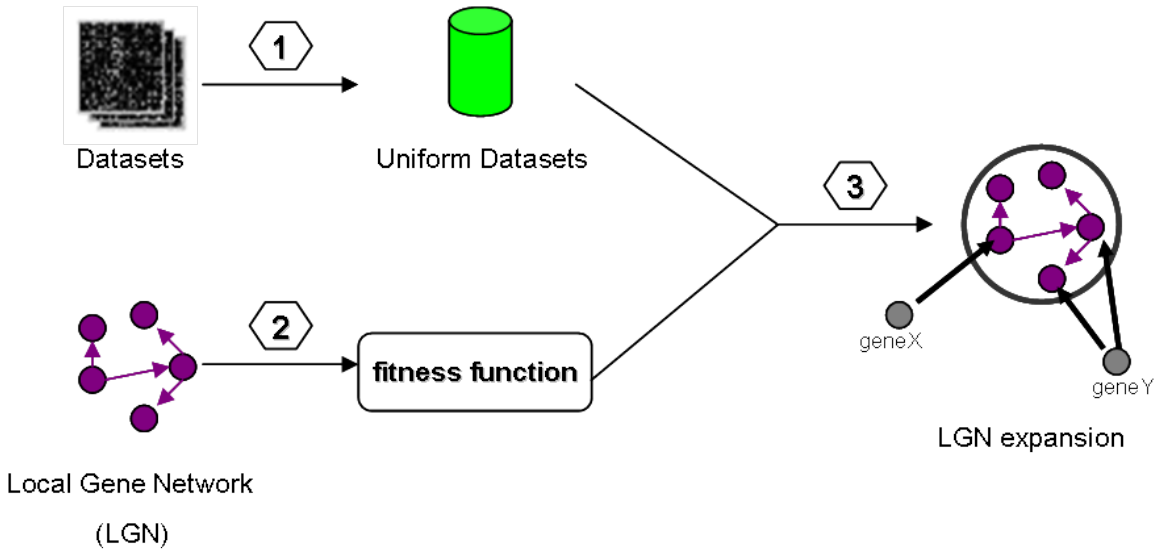


Figure 2.9: **Overview of the GENESYS algorithm.**

the variables is described by the function $f^v : \mathbb{C}^{|U|} \rightarrow \mathbb{C}$ for each $v \in \mathbb{U}$ (the value of v at time t depends on the values of its input variables at time (t-1)). The prior knowledge is used at the beginning to describe a model space. This is defined by the quadruple ($\mathbb{U}$, $\mathbb{C}$, F_{bio}, G_{bio}), where $\mathbb{U}$ and $\mathbb{C}$ are the sets defined above, F_{bio} is the class of the candidate f^v and G_{bio} is a class of dependency graphs on $\mathbb{U}$. F_{bio} and G_{bio} are used to limit the model space and incorporate the prior knowledge of the LGN.

Fitness evaluation is a critical step to LGN expansion. The fitness function uses the idea presented in Liang et al. [1998] and it must perform well in term of sensitivity, precision and computing efficiency. In GENESYS there are two types of fitness: local and global. The local fitness function evaluates the fitness of the experimental data to the function f^v of a single variables v, while the global function evaluates the overall network. Summarizing GENESYS starts from the LGN (G') and outputs G", namely the LGN expansion (G'$\subseteq$G"). The fitness value of G' is determined and then one gene at a time ($v \in \mathbb{U}$) is added to G' and the new fitness is calculated. Only the genes that have an improvement of fitness respect to the raw G' are selected and included in the G" [Tanay et al., 2001].

Hashimoto et al. [2004] developed the **Growing algorithm** that uses gene expression data to discover subnetworks of a large network, in which genes must to have two principal characteristics:

- genes of the subnetwork must be significantly related between them;

- genes of the subnetwork must be not strongly conditioned by genes outside the

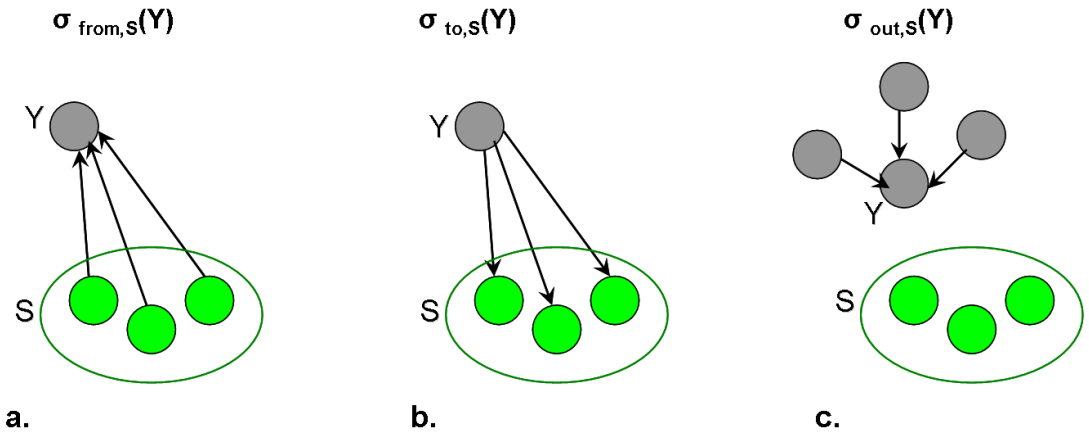


Figure 2.10: **Parameters generated by Growing algorithm**
S is a gene of a LGN and $\mathbb{Y}$ is a set of all genes excluded S. Figure 2.10-a. represents the **from** impact of $\mathbb{Y}$ to S. Figure 2.10-b. presents the depiction of the **to** impact of $\mathbb{Y}$ to S. Figure 2.10-c. shows the measure of the strength of edge from genes external to S to $\mathbb{Y}$ (adapted from Hashimoto et al. [2004]).

subnetwork

In particular this method starts from a little initial group of genes ('seed') and then adds new genes expanding the seed in a greater subnetwork. To reach this aim, the Growing algorithm proceeds modeling a GRN as a directed graph in which at each relationship between the variables is associated a coefficient of determination. "The coefficient of determination measures the degree to which a set of variables improves the prediction of a target variable relative to the best prediction in the absence of any conditioning observations"[Hashimoto et al., 2004].This means that the influence is used to measure the strength of a relationship and with the term $\sigma_X(Y)$ is indicated the sum of influences of the genes in X on the set of genes Y. In particular if S is a a gene of a LGN and $\mathbb{Y}$ is a set of all genes excluded S then are measured three coefficient of determination:

- $\sigma_{from,S}$(Y): the collective strength of connection from the to the target set of genes $\mathbb{Y}$;
- $\sigma_{to,S}$(Y): the impact (strength of connection) of $\mathbb{Y}$ to S;
- $\sigma_{out,S}$(Y): the measure of the strength of edge from genes external to S to $\mathbb{Y}$

Figure 2.10 represents the three determination coefficients that are computed from Growing algorithm. Once computed $\sigma_{from,S}$(Y), $\sigma_{to,S}$(Y) and $\sigma_{out,S}$(Y) the algorithm combines

them in a unique measure that is the final strength measure of the gene selection. In fact are selected those genes of the subnetwork that improve the value of the strength with respect to the values of genes of the LGN.

Gat-Viks and Ron Shami system. In 2007 Gat-Viks and Ron Shamir developed a system (Figure 2.11) that, starting from prior knowledge of a LGN, adds interactions between the genes of the LGN and it expands the LGN with additional genes and relative relationships [Gat-Viks and Shamir, 2007]. It starts from prior biological knowledge and it formalizes this in the Bayesian network in which each variable (node) can have several discrete states and then it obtains the Bayesian scoring matrix. The method finds the discrete function which represents the different relationships between genes of the LGN. The next step consists in generating two evaluation models starting from different levels of expression: observed and predicted. The observed expression level derives from a measurement in biological experiments (gene expression data, measures of the metabolism and/or, proteins). The predicted expression level, instead, is the probabilistic expectation of the variable given the model and the experimental data (gene expression data of the genetic perturbation). In the final steps these two expression levels are compared and the disagreement, between observed and predicted expression levels, indicates the possible edge to be added on the LGN. The new score of the Bayesian matrix with these new edges is calculated and if its score is bigger of the score of the original model the edge is added to the LGN. The same method is used to expand a LGN. Each hypothetical expansion gene is added to the LGN and the new scoring matrix is recalculated [Gat-Viks and Shamir, 2007].

BN+1 [Hodges et al., 2010] (Figure 2.12) is an expansion algorithm created to discover genes, not included in the LGN, that generates the best network score when a gene is added to an existing core network topology. This system uses prior knowledge of the LGN and gene expression data as the starting point to generate a bayesian network, termed core BN. The criterion used to arrive to this core BN is the log posterior score. BN+1 starts from the known genes (genes of the LGN) and uses the independent simulation to generate randomly networks with these genes. Each network is scored using log of the Bayesian Dirichlet metric (BDe) [Heckerman et al., 1995] and the posterior distribution is estimated. Finally a consensus network, also called core network, is generated from the random networks considering the best log posterior score. At this step, for the core network, the direction is determined: the directional edges represent those edges that appear to have the same direction in all the random networks. Instead the undirected relationships are the edges that appear in all the random networks, but which different directions. The expansion step adds a gene at a time to the core network and the new scores are computed. The gene added is present in the expression data used in input data.

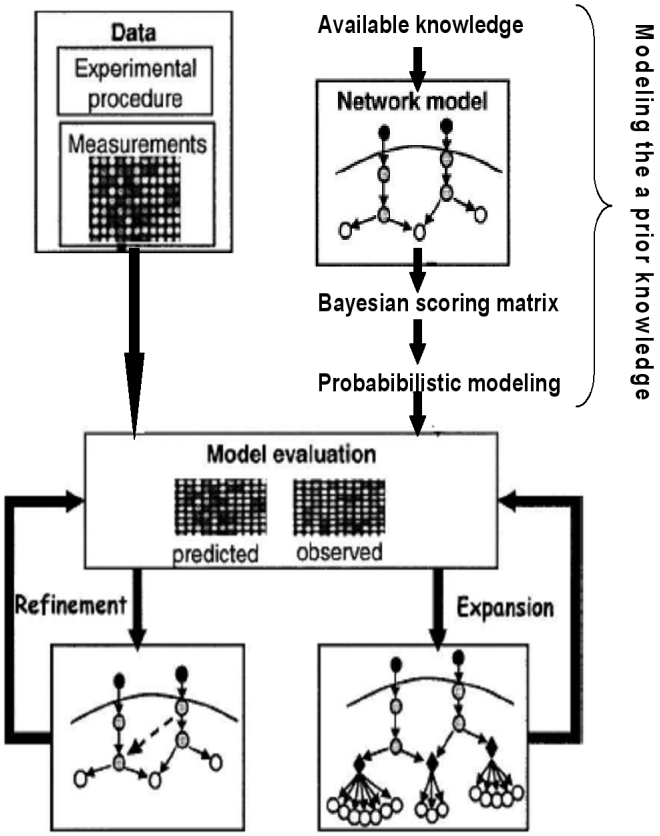


Figure 2.11: **Representation of the Gat-Viks and Ron Shamir methodology** (adapted from Gat-Viks and Shamir [2007]).

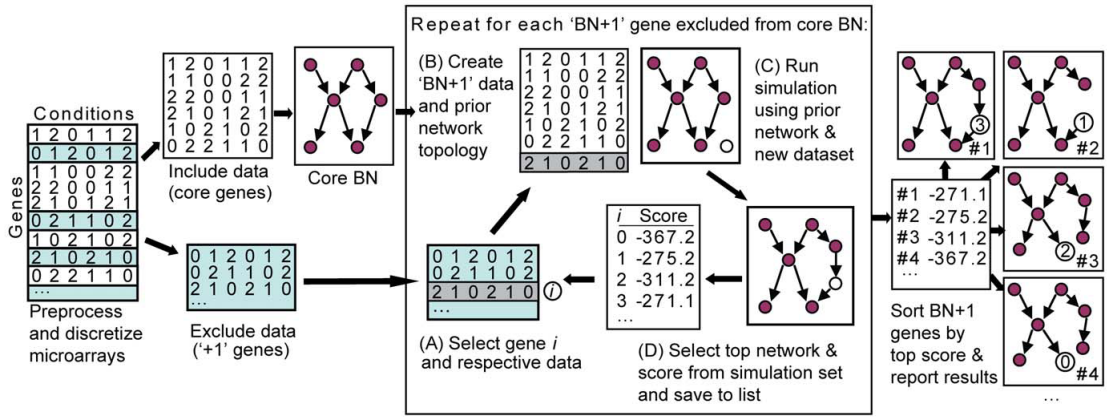


Figure 2.12: **Schematic representation of the BN+1 expansion algorithm** derived from Hodges et al. [2010].

At the end, the genes that expand the LGN will be those that improve the core network score.

GENIES [Kotera et al., 2012] discovers the new genes related with a specific LGN using different type of data in combination or alone. These data are: gene expression data, protein localization data, phylogenetic profile, kernel matrix based on the gene expression profile, kernel matrix based on the protein localization profile and kernel matrix based on the phylogenetic profile. Initially this method use a kernel function to transform data sets in a kernel similarity matrix (e.g. correlation coefficient matrix), where each element corresponds to a gene-gene similarity [Kotera et al., 2012]. Subsequently, GENIES proceeds with a training process in which the elements of the LGN (genes or proteins) are mapped in a feature space, where interacting elements are close to each other and the Euclidean distance is calculated [Yamanishi et al., 2005]. Euclidean distance is considered to be the indicator of the presence/absence of edges and it will set the threshold. After the training process the next step is the test process for the testing. In this phase, other genes (not included in the LGN) are mapped in the Feature space and only genes that have a Euclidean distance above the threshold will be in the final Gene Network [Yamanishi et al., 2005]. Figure 2.13 is a schematic drawing of all phases of GENIES.

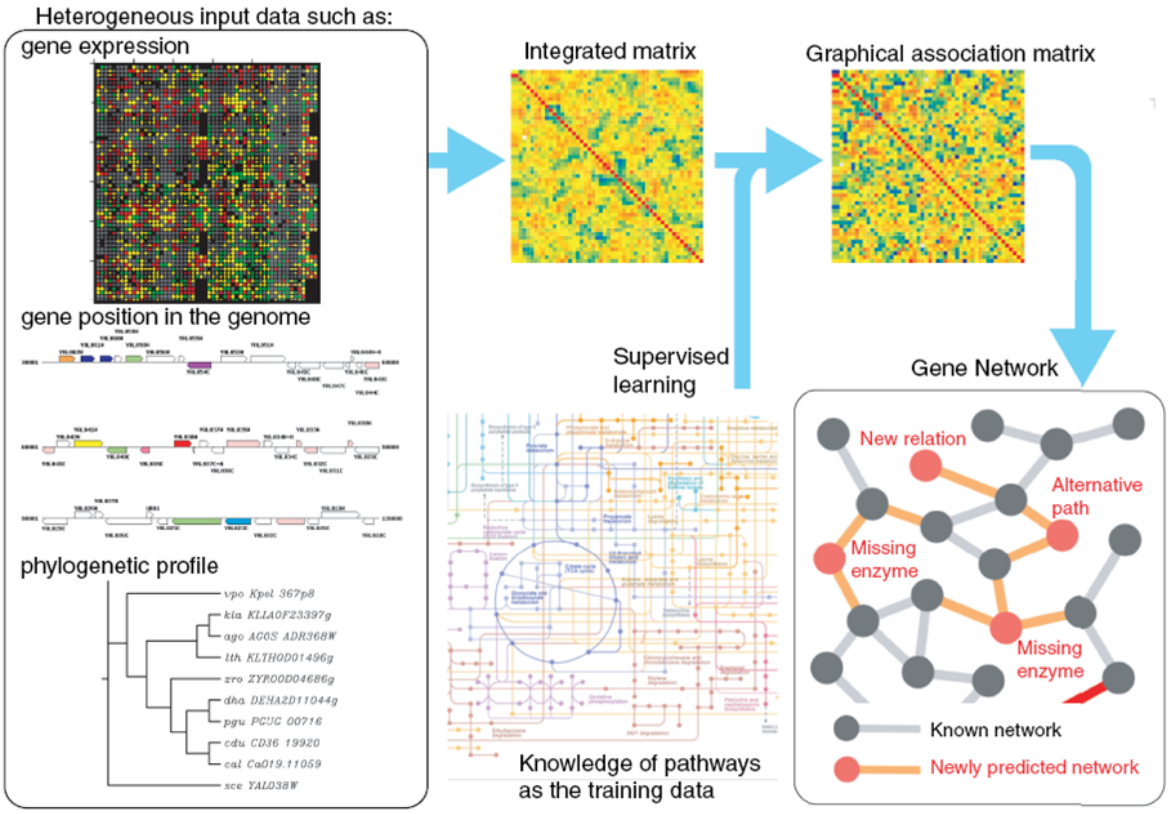


Figure 2.13: **Overview of GENIES**(taken from Kotera et al. [2012].

Chapter 3

PC-Iterative Method (PC-IM)

The PC-Iterative method (PC-IM) has been developed during this thesis work to expand a known Local Gene Network (LGN) taking advantage of an algorithm already existing in literature. The expansion involves the discovery of genes related to the genes of the LGN. These new genes are found from the analysis of thousands of genes. For example PC-IM can analyse all the genes of a gene expression experiment and/or of a whole genome. The analysis foresees to use the gene expression data of all input genes (LGN-genes and other genes) and takes advantage from the algorithm capacity of infering causal relationships between LGN-genes and other genes. Finally PC-IM gives a list of genes that expands the known LGN.

The algorithm used in PC-IM is the PC algorithm, specifically the PC algorithm implementation included in the R-package pcalg (R-package pcalg: `//www.r-project.org`) [Kalisch et al., 2010]. One problem of the PC algorithm is the possibility to analyse together only a limited number of variables (1000 genes maximum to have a good efficiency) [Wang et al., 2010]. For this reason our method divides the input gene-set in subsets (*tiles*). All the *tiles* have the same size and all of them contain the genes of the LGN.

Another important feature of PC-IM is the intrinsic evaluation of its performances. This property allows an estimation of the precision of the final expansion gene list of the LGN. The intrinsic evaluation is computed in terms of the following measures:

$$PPV = \frac{TP}{(TP + FP)} \tag{3.1}$$

$$Se = \frac{TP}{(TP + FN)} \tag{3.2}$$

$$1 - Sp = \frac{FP}{(FP + TN)} \tag{3.3}$$

Where TP, FP, TN and FN are the number of true positive, false positives, true negatives and false negatives, respectively. The TP is the number of genes correctly predicted by PC-IM, FP is the number of incorrectly predicted genes, TN is the number of correctly identified genes that are not involved in the expansion of LGN and FN is the number of true expansion genes missed from the algorithm. Se is the Sensitivity (or Recall) and it represents the ability to retrieve in the prediction an edge or a node when this is present in the real network. PPV indicates the Positive Predicted Value (or Precision), namely the accuracy on the inferred network. 1-Sp is the false positive rate (FPR) and it refers to the rate of genuine negatives considered to be positives. The parameters Se and 1-Sp are used to plot the Receiver Operating Characteristic curves (ROC) and the area under this curve (AUC) is calculated. In addition to the ROC curve, the Precision versus Recall (PR) curve is determined together with the minimum distance (d_{min}) between each point of this curve from the point of (1,1) coordinates. At the end of the process, to establish the final list of the expansion genes, PC-IM will make the best compromise between values of the AUC and d_{min}. PC-IM can be divided in five steps as represented in Figure 3.1.

The five steps are:

Step 1: *tiles* creation;

Step 2: run of the PC algorithm;

Step 3: frequencies computation;

Step 4: intrinsic performances assessment;

Step 5: cut-off frequency application.

Before describing the steps we list data and parameters necessary to run PC-IM:
Input data:

kngenes are the genes of LGN that will be expanded;

knedges are the relationships between genes of the LGN that will be expanded (optional input data);

obs-gene are expression data. A nxm gene expression matrix containing n genes in m cases.

Parameters:

t is the size of the *tiles*, namely the number of genes in each *tile*;

i is the number of iterations. This number specifies the times the whole genome is divided into *tiles*;

D is the number of sub-networks of LGN (subLGN);

d is the size of the subLGNs.

Output:

Expansion-genes is a list of genes that expands the LGN with their relative frequency.

Step 1: *Tiles* creation

The genes of the genome are randomly divided in non overlapping *tiles* of size t. Each *tile* is merged with the set of genes of the input LGN. This operation is repeated i times. Adding the LGN genes to each *tile* permits to potentially infer the causal relationships of these genes with the other genes of the genome.

Step 2: run of the PC algorithm

PC runs on each *tile* using the gene expression data. As a result it gives a network (nodes and relationships) for each run. From these networks are extracted the sub-networks that include both the genes and relationships within the LGN genes and between LGN genes and the external ones, namely genes in the original *tile* and consequently not included in the LGN.

Step 3: frequencies computation

PC-IM creates a unique list of genes called expansion list. This list contains all the genes present in at least one single sub-network. For each gene in the expansion list, the algorithm calculates the frequency. This absolute frequency of a gene is the number of times that the gene is present in the sub-networks. PC-IM computes both the absolute and the normalized frequency. The latter is obtained dividing the absolute frequency of a gene with the number of times that the same gene could be found, that is how many times it was present in the *tiles*. Usually this number coincides with the number of iterations. For example if i=100 and x gene has 90 as absolute frequency, then the normalized frequency will be 0.9.

Step 4: intrinsic performances assessment

In this step PC-IM assesses its own performance and establish the minimum normalized frequency value necessary to have the best expansion performance. The assessment is done using the information of the original LGN. Initially from the LGN are extracted D subLGNs of size d. With this operation the genes of the expansion list are split in three different categories:

subLGN genes: genes of the LGN;

INTRA genes: genes of the LGN, that are not included in the subLGN;

EXTRA genes: the additional genes randomly selected from the genome.

Now the evaluation parameters are calculated considering the expansion of each subLGN. A gene in the list will be considered to be positive if it is among the other genes of the LGN that are let in the subLGN (intra gene) and negative if it is among the genes (extra gene). This operation is repeated for each of the D subLGNs and for each relevant frequency value. In practice, we use as Gold Standard the information on the LGN given in the input data. The evaluation measures are Positive Predictive Value (PPV or Precision), Sensitivity (Se) and False Positive Rate (FPR). Mathematically, they are defined by:

$$PPV = \frac{TP_{intra}}{(TP_{intra} + FP_{extra})} \tag{3.4}$$

$$Se = \frac{TP_{intra}}{(TP_{intra} + FN_{intra})} \tag{3.5}$$

$$FPR = \frac{FP_{extra}}{(FP_{extra} + (TN_{intra} + TN_{extra}))} \tag{3.6}$$

where TP, FP, FN and TN are the number of the true positive, false positive, false negative and true negative respectively. The terms intra and extra have the same meaning specified above. The Se and FPR with different frequencies are used to plot ROC curve and the area under ROC curve (AUC). PPV and Se values are used to plot the Precision-Recall (PR) curve and the minimum distance (d_{min}) from point (1,1) is calculated. This step returns the frequency that gives the maximum value of AUC and the minimum value of d_{min} or that gives the best compromise between the higher AUC value and the smaller d_{min} value.

Step 5: cut-off frequency applications In this last step PC-IM determines which genes, in the gene expansion list, will be returned as the final output of the method. The selection applies the cut-off frequency computed at Step 4 as a cut-off frequency on the gene expansion list computed at Step 3. The final list of genes is returned. The cut-off frequency is the minimum frequency that the genes, in the gene expansion list, must have in order to be considered involved in the expansion of the LGN.

PC-IM pseudocode is reported below:

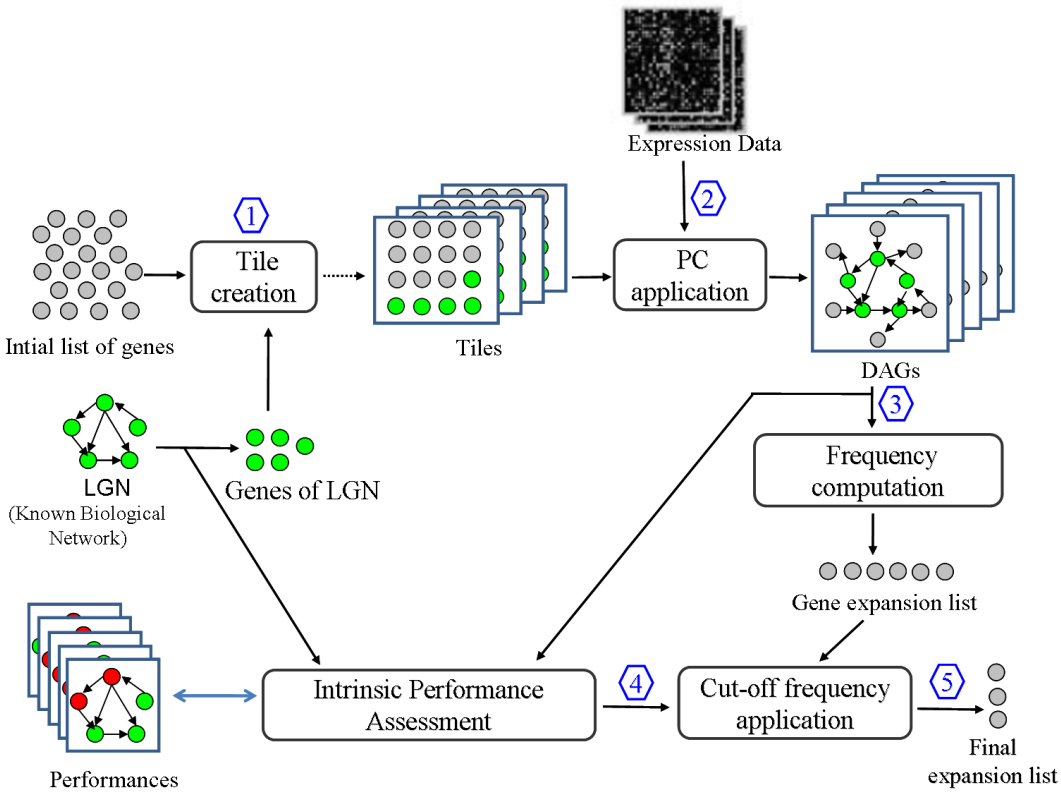


Figure 3.1: **Schematic representation of PC-IM.**
Numbers into the hexagonal boxes are the ordered steps of this method. **1** is the "*Tiles* creation", **2** is "run of the PC algorithm", **3** is "frequencies computation", **4** is "intrinsic performances assessment"and **5** is "cut-off frequency applications".

Input data:

```
kngenes := genes of the LGN that will be expanded
knedges := relationships between genes of the LGN that will be expanded
obs_data := n x m expression matrix (n nodes and m cases)
```

Parameters:

```
t := tiles size
i := number of iterations
D := number of sub-networks of LGN (subLGN)
d := subLGNs size
```

Output:

```
Expansion_genes := list of genes that expands the LGN with their relative frequency
```

Step1: *Tiles* creation

```
n_num:=rownames of obs_data (initial set of genes of cardinality n)
n_lgn:= genes number of the kngenes
genes⊆{n_num \ kngenes}
N_sub := (length(n)-n_lgn)/t
p := i*N_sub
create a list R[p]
create a list pgenes[N_sub]
l:= 1
z:= 1
while (l<i) {
  while(z<N_sub){
      pgenes[z] := randomly choose (n-n_lgn) elements from genes and add
          kngenes
      z:= z+1
        }
  R[l] := pgenes[N_sub]
  l:= l+1
    }
```

Step2: run of the PC algorithm:

```
create a list pcpred[p]
l:=1
while (l<p) {
  exp_data⊂obs_data, where
                exp_data r x m , r < n and r = |R[l]|
```

```
                exp_data = rows(exp_data= R[l]) Λ (columns (exp_data) =
                   columns(obs_data))
  run PC algorithm on exp_data
  pcpred[l]:= results of PC run
      }
create a gene network M in which there are all the results of the pcpred
```

Step3: frequencies computation:

```
G is a list of all genes in the M gene networks that have a
   relationships with kngenes
i:=1
while (z<|G|) {
  F_G := number of times in which the gene in position G[z] appears in
     the R list;
  f_G := number of times in which the gene position G[z] appears in the
     M gene network;
  freq_ratio  := f_G / F_G
  add a G[z] the freq_ratio
  z := z+1
      }
order G respect to the freq_ratio
```

Step4: intrinsic performances assessment:

```
create a list subLGN[D]
l := 1
while (l<D) {
  subLGN[l] = randomly choose d elements from kngenes
  l := l+1
      }
delta := (max(G[freq_ratio])-min(G[freq_ratio))/100
freq_%  := (G[freq_ratio]-max(G[freq_ratio])+100*(delta))/delta
create a list New_freq
New_freq = add freq_% to G
rls_intraLGN  := relationships among genes of the LGN
while (l<100) {
  New_freq[freq_%] = [l]
  z := 1
  while (z<D) {
    TP_FN[z] = {kngenes\subLGN[z]}∩{subLGN[z]∩rls_intraLGN}
    TP_FP[z]⊂G that excludes the genes without relationships with genes
       of subLGN[z] and also genes of the subLGN[z] that have a
       relationships with other genes of the subLGN[z]
    TP[z] = TP_FN[z]∩TP_FP[z]
```

```
    TN[z] = {obs_gene[n]\{subLGN[z]∪TP_FN[z]}}
    z := z+1
        }
  TP_FP_mean  := mean of TP_FN[l]
  TP_FP_mean  := mean of TP_FP[l]
  TN_mean   := mean of TN[l]
  New_Freq[l] := add {TP_FN[l]∩TP_FP[l]∩TN[l]} to New_Freq[l]
  l := l+1
      }
compute ROC curve on New_Freq
d(ROC) := distance of point [0, 1] of ROC curve
compute PR curve on New_Freq
d(PR) := distance of the point [1, 1] of PR curve
cut-off := value of New_Freq[freq_%] which correspond to min(d(PR))
```

Step5: cut-off frequency applications:

```
ntw_expd⊆{G \ kngenes}
create list Expansion_genes
delta := (max(ntw_expd[freq_ratio])-min(ntw_expd[freq_ratio]))/100
freq_% := (ntw_expd[freq_ratio]-max[ntw_exp[freq_ratio])+100*delta)/
   delta
add freq_% to ntw_expd
Expansion_genes⊂ntw_expd(ntw_expf[freq_%]≥cut-off)
```

The difference between our method and the other expansion algorithms are described in the Section 2.3 and are schematized in Table 3.1.

	PC-IM	**GENESYS**	**Growing-algorithm**	**Viks-and-Shamir-algorithm**	**BN+1**	**GENIES**
Authors	Coller et al. (European Patent application EP13151728.6 (of date 17 January 2013))	[Tanay et al., 2001]	[Hashimoto et al., 2004]	[Gat-Viks and Shamir, 2007]	[Hodges et al., 2010]	[Kotera et al., 2012]
Prior knowledge of the LGN	start and final step	start point	start point	start point	start point	start point
Type of input data	gene expression data	gene expression data	gene expression data	different type of data	gene expression data	different type of data
Analyzed genes	all tiles genes at a time	one at a time	one at a time	one at a time	one at a time	one at a time
Selection criterion for expansion genes	Frequency	Fitness function	Coefficient of determination	Bayesian score	Log BDe	Euclidean distance

Table 3.1: **Comparison of different LGN expansion algorithms.**

Chapter 4

Evaluation of the PC-Iterative Method (PC-IM)

This chapter describes both the procedure, used to analyse the PC-IM method, as well as the results of this evaluation.

4.1 Preliminary evaluation 1: *in silico vs in vivo*

Once a new algorithm or a new method is developed is necessary to test its performance (in terms of PPV and Se) and to compare it with those of the other algorithms proposed by the literature. To test and compare algorithms is possible to choose between two type of data: *in vivo* or *in silico* data. The aim of this preliminary evaluation is to understand which of these two type of data better suits the evaluation of PC-IM. The type of the gene expression data is chosen comparing the PC algorithm performances versus ARACNE algorithm performances on *in silico* data (DREAM) *versus in vivo* data (M3D and GEO). To weight the difference of the inference GRN with *in silico* from *in vivo* data we have compare Positive Predictive Value (PPV or Precision) (Formula 3.1) and Sensitivity (Se or Recall) (Formula 3.2).

4.1.1 *In silico* data

In silico data derive from mathematical equations and simulate the value of *in vivo* data. This type of data may be generated by the programmer of the algorithm for its testing or can be downloaded from free websites.

The Dialogue on Reverse Engineering Assessment and Methods (DREAM) project [Stolovitzky et al., 2007] [Stolovitzky et al., 2009] is an example of *in silico* data that can be freely downloadable from the website (`http://wiki.c2b2.columbia.edu/dream/index.`

`php/The_DREAM_Project`). DREAM data are divided in Challenges, each one represents a different biological task (e.g. inference of gene networks, prediction of protein-protein interactions). Moreover in each Challenge there are *in silico* data and their relative Gold Standards (GS) [Stolovitzky et al., 2007].

In this section we chose the Challenge about inference of gene networks, in particular we have used Challenge 2 of DREAM4: In Silico Network Challenge [Marbach et al., 2009]. This Challenge is divided in three sub-challenges that differ in the size of the network. Each sub-challenge contains five types of datasets (wild-type, knockout, knockdowns, multifactorial perturbations and time-series) and five GS networks (ones for each type of the datasets). The GS networks are gene networks that have been described in same organisms and are considered therefore the Gold truth in terms of relationships between the network elements (genes). Each dataset present the simulated data for two organisms: *Escherichia coli* and *Saccharomyces cerevisiae*.

The **sub-challenges** are three (InSilico-Size10, In Silico-Size100 and InSilico-Size100-Multifactorial) and are formed in this way:

- **InSilico-Size10**: it contains five networks of 10 genes (size 10);

- **InSilico-Size100**: it contains five networks of 100 genes. In this sub-challenge the multifactorial perturbation datasets are not included as they are the subject of another sub-challenge (InSilico-Size100-Multifactorial);

- **InSilico-Size100-Multifactorial**: it contains five networks of 100 genes (as InSilico-Size100), but it presents only the multifactorial dataset;

The **datasets** are:

- **Wild-type**: *in silico* data of the unperturbed network;

- **Knockout**: *in silico* data in which each k-th data line is the steady-state of the network after knockout (deletion) of gene k. An independent deletion is present for the all the k genes of the network;

- **Knockdowns**: *in silico* data where a knockdown of every gene of the network is simulated. The expression values are obtained with reducing by half the transcription rate of the corresponding gene;

- **Multifactorial perturbations**: *in silico* data which correspond to the steady-state levels of variations of the network, obtained by applying multifactorial perturbations to the original network;

- **Time-series**: time course data of the network changes after a perturbation. The perturbation (strong increase or decrease of the gene basal activation) do not regard all genes, like in multifactorial perturbation, but only a third of them. For networks of size 10 we considered 5 different time series, for networks of size 100 we considered 10 time series. Each time series comprises 21 time points.

From all these available sub-challenges and datasets we have used only the sub-challenges InSilico-Size10 and InSilico-Size100 and the Time-series dataset. The names of the different Gene Regulatory Networks (GRNs) tested are summarized in Table 4.1 where rep1 and rep2 are the names for the GRN of the *Escherichia coli*. rep3, rep4 and rep5 are the GRN-name for the *Saccharomyces cerevisiae*.

Size GRN (nr genes)	Escherichia coli	Saccharomyces cerevisiae
10	rep1; rep2	rep3; rep4; rep5
100	rep1; rep2	rep3; rep4; rep5

Table 4.1: **Description of the DREAM4-Challenge 2 (time series) GRN.**
The GRN name identifies only the organism and not the size of GRN.

In Table 4.2 and Table 4.3 the results given by the two tested algorithms (PC and ARACNE) with *in silico* gene expression data DREAM 4, Challenge 2 (time series) are reported. In the case of the regulatory network of 10 genes, ARACNE shows a PPV greater than the PC algorithm (rep1 and rep2 in Table 4.2; rep3, rep4 and rep5 in Table 4.3). In the case of the transcriptional regulatory network with 100 genes ARACNE has only a slightly better PPV. Looking at Sensitivity, the PC algorithm is in most trials better than ARACNE and this is more evident with a small number of genes (rep1 and rep2 in Table 4.2; rep 3, rep4 and rep5 in Table 4.3).

4.1.2 *In vivo* data

In vivo data derive from real biological experiments (e.g. the *in vivo* data in case of expression data are usually from microarray hybridizations) and they are freely available in specialized repositories. Since the GS DREAM data derive from real biological networks we have chosen to compare results from the *in silico* data of *Saccharomyces cerevisiae* with the *in vivo* data of this organism. In particular we have used two different types of *in vivo* data, to test if there exist differences, in the algorithms' output, according to the type of *in vivo* gene expression data. In the first analysis we have considered the networks GRN rep3, rep4 and rep5 of DREAM 4 Challenge 2 and the expression data available on the m3d (`http://m3d.bu.edu/norm/`) [Faith et al., 2008] database. The m3d, is a

size of GRN	algorithm	TP	FP	FN	PPV	Se
10 (rep1)	PC	10	5	5	0.667	**0.667**
10 (rep1)	ARACNEC	3	0	12	**1.000**	0.200
10 (rep2)	PC	5	8	11	0.385	**0.313**
10 (rep2)	ARACNE	3	1	13	**0.750**	0.188
100 (rep1)	PC	28	156	148	0.152	**0.159**
100 (rep1)	ARACNE	23	116	153	**0.165**	0.131
100 (rep2)	PC	14	156	235	**0.082**	**0.056**
100 (rep2)	ARACNE	10	116	239	0.079	0.040

Table 4.2: **Description of DREAM4-Challenge 2 (time series data of *Escherichia coli*).** DREAM4, Challenge 2, time series, *Escherichia coli* transcriptional regulatory network (rep1 and rep2) and two different sample size with 10 genes (size 10) and 100 genes (size 100).

size of GRN	algorithm	TP	FP	FN	PPV	Se
10 (rep3)	PC	6	5	9	0.545	**0.400**
10 (rep3)	ARACNE	4	1	11	**0.800**	0.267
10 (rep4)	PC	7	5	6	0.583	**0.538**
10 (rep4)	ARACNE	5	0	8	**1.000**	0.385
10 (rep5)	PC	10	5	2	**0.667**	**0.833**
10 (rep5)	ARACNE	7	0	5	0.538	0.583
100 (rep3)	PC	36	134	159	0.212	0.185
100 (rep3)	ARACNE	37	89	158	**0.294**	**0.190**
100 (rep4)	PC	30	163	181	0.155	**0.142**
100 (rep4)	ARACNE	29	109	182	**0.210**	0.137
100 (rep5)	PC	35	155	158	0.184	**0.181**
100 (rep5)	ARACNE	33	100	160	**0.248**	0.171

Table 4.3: **Description of DREAM4-Challenge 2 (time series data of *Saccharomyces cerevisiae*).**
DREAM4, Challenge 2, time series, *Saccharomyces cerevisiae* transcriptional regulatory network (rep3, rep4 and rep5) and two different sample size with 10 genes (size 10) and 100 genes (size 100).

specific repository for *Saccharomyces cerevisiae*, containing 904 array experiments. In the second case we analysed only GRN rep3 of DREAM 4 Challenge 2. The size of this LGN is 10 genes and the type of relationships and regulatory interaction between the LGN is

reported in Table 4.4. Gene expression data were selected from the GEO Database (`http://www.ncbi.nlm.nih.gov/gds?term=saccharomyces`). We have selected only those gene expression data that appeared to be inherent with the specific GRN (Table 4.5).

index	gene	index	gene	type regulation
G1	YBR182C	G2	YHL027W	+
G2	YHL027W	G4	YGR249W	-
G2	YHL027W	G6	YPR065W	-
G3	YGL162W	G2	YHL027W	-
G3	YGL162W	G4	YGR249W	-
G3	YGL162W	G5	YPL177C	-
G5	YPL177C	G7	YIL101C	-
G6	YPR065W	G3	YGL162W	+
G6	YPR065W	G7	YIL101C	-
G7	YIL101C	G1	YBR182C	-
G7	YIL101C	G4	YGR249W	+
G7	YIL101C	G8	YJR147W	+
G8	YJR147W	G4	YGR249W	+
G9	YOR113W	G7	YIL101C	+
G10	YOR363C	G7	YIL101C	+

Table 4.4: **Description of the 10 genes of the GRN 3, DREAM4-Challenge 2 and relative type of interactions among these genes.**

In the first analysis, with m3d gene expression data, the sensitivity of PC appared to be than that of ARACNE with 10 genes in the GRN, instead with 100 genes sensitivity values between PC and ARACNE are comparable (Table 4.6). For both algorithms the values of the sensitivity are lower than those obtained with *in silico* data. In fact when the size of GRN is 10 genes, PC algorithm sensitivity ranges between (0.231-0.400) with *in vivo* data and between (0.400-0.833) with *in silico* data, while for ARACNE ranges between (0.077-0.250) with *in vivo* data and between (0.267-0.583) with *in silico* data. In case of 100 genes in the GRN, PC ranges between (0.005-0.072) for the *in vivo* data and between (0.142-0.185) for the *in silico* data. ARACNE ranges between (0.005-0.081) and (0.137-0.190) for the *in vivo* and *in silico* data respectively. The value of PPV with *in vivo* data is also smaller that with *in silico* data for both algorithms and also in this case ARACNE shows the best value of PPV. The range values of PPV with 10 genes in the GRN are (0.214-0.429) and (0.545-0.677) for the PC algorithm with *in vivo* and *in silico* data respectively. Instead for ARACNE they are (0.500-0.600) with *in vivo* data

GEO Code	type of experiments	number of experiements
GDS112	HS	15
GDS36	HS	8
GDS34	HD	10
GDS16	HS	36
GDS2343	HS	8
GDS1711	HS	12
GDS33	osmotic	10
GDS20	osmotic	12
GDS2002	anaerobic	30
GDS3030	anaerobic	6

Table 4.5: **Description of the gene expression data from GEO database.**
The GEO database represented are inherent with the genes present in the Table 4.4. (HS is Heat Shock gene observational expression data; osmotic means osmotic gene observational expression data; Anaerobic is Anaerobic gene observational expression data)

and (0.583-1.00) with *in silico* data. When the GRN has 100 genes PPV ranges observed for the PC are (0.006-0.990) and (0.155-0.212) and for ARACNE they are (0.007-0.115) and (0.210-0.294) with *in vivo* and *in silico* data respectively.

In the second analysis that uses gene expression data related to the genes of the GRN, we have observed a change in PPV and Se behaviour. This is evident in Table 4.7. In this case the value of PPV of ARACNE and PC are comparable but in one case, with osmotic gene expression data, PC has greater PPV than ARACNE. Moreover the range of the PC algorithm PPV values has improved with respect to *in silico* data, (0.308-0.800) for GEO dataset (*in vivo* data) *versus* (0.545-0.677) *in silico* data and 10 genes in the GRN. Instead this range value has worsened for ARACNE, (0.500-n.d.) *versus* (0.583-1.00), *in vivo versus in silico* data (n.d. means not determined). For the sensitivity, PC algorithm showed the best performances and, similarly to what observed with the m3d gene expression data, the sensitivity values are decreasing with this type of data for both algorithms. The PC ranges are between (0.267-0.333) for *in vivo* data (GEO Dataset) and (0.400-0.833) with *in silico* data. For ARACNE the ranges are between (0.067-0.133) and (0.267-0.583) for *in vivo* and *in silico* data respectively.

4.1.3 Discussion of the results of preliminary evaluation 1

In the preliminary evaluation 1 we have tested the two algorithms PC and ARACNE in their ability to infer GRNs starting with two different type of gene expression data:

size of GRN	algorithm	TP	FP	FN	PPV	Se
10 (rep3)	PC	6	8	9	0.429	**0.400**
10 (rep3)	ARACNE	3	2	12	**0.600**	0.200
10 (rep4)	PC	3	11	10	0.214	**0.231**
10 (rep4)	ARACNE	1	1	12	**0.500**	0.077
10 (rep5)	PC	3	12	8	0.250	**0.333**
10 (rep5)	ARACNE	3	2	9	**0.600**	0.250
100 (rep3)	PC	14	128	181	**0.099**	**0.072**
100 (rep3)	ARACNE	12	164	183	0.068	0.062
100 (rep4)	PC	13	143	198	0.083	0.062
100 (rep4)	ARACNE	17	131	194	**0.115**	**0.081**
100 (rep5)	PC	1	155	192	0.006	0.005
100 (rep5)	ARACNE	1	147	192	0.007	0.005

Table 4.6: **DREAM4-Challenge 2, *Saccharomyces cerevisiae.*** *In vivo* data (m3d gene expression data). Number of Experiments: 904.

in silico and *in vivo*. The aim of this test was to understand which is the type of data were more useful to explore the properties of PC-IM. The results show that the values of the performances (PPV and Se) change appreciably between *in silico* and *in vivo* data and also depending on the different *in vivo* data. This underlines how the method of generation of *in silico* data influences the comparison between the algorithms and how it can introduce biases in judging an algorithm's performances.

In vivo data have the advantage that they are obtained from real hybridization experiments and this overcomes the possibility to advantage the performances of an algorithm by creating *ad hoc* datasets. However *in vivo* data have the problem of the absence of a Gold Standard. In fact, when we infer the GRN from *in vivo* data, there are not true outputs to evaluate the algorithm's performances against. Nevertheless to validate PC-IM, in the expansion task (and not inference) of a GRN, called Local Gene Network (LGN), we chose the *in vivo* data.

4.2 Preliminary evaluation 2: PC algorithm *versus* ARACNE algorithm performing LGN expansion

In the preliminary evaluation 2 we compared the performances of PC and ARACNE on a real LGN expansion task using *in vivo* data. In step 2 of PC-IM is possible to choose any algorithm to expand the LGN. With this test we show why we use the PC algorithm.

size	algorithm	TP	FP	FN	PPV	Se
HS	PC	4	9	11	0.308	**0.267**
HS	ARACNE	2	3	13	**0.400**	0.133
osmotic	PC	4	1	11	**0.800**	**0.267**
osmotic	ARACNE	1	1	14	0.500	0.067
anaerobic	PC	4	5	11	**0.444**	**0.268**
anaerobic	ARACNE	0	0	15	n.d.	0.000
HS+osmotic+anaerobic	PC	5	9	10	0.357	**0.333**
HS+osmotic+anaerobic	ARACNE	2	3	13	**0.400**	0.133

Table 4.7: **DREAM4-Challenge 2, *Saccharomyces cerevisiae*, rep3, size 10. *In vivo* data (GEO).** (HS is Heat Shock gene observational expression data; osmotic means osmotic gene observational expression data; anaerobic is anaerobic gene observational expression data; HS+ osmotic+ Anaerobic is sum of the gene observational data submitted before, n.d. means not determine)

The choice of these two algorithms was made for the following reasons:

- PC is an algorithm developed in the social science field and its application to gene network inference starting from gene expression data is quite recent [Spirtes and Glymour, 1991];
- ARACNE can be applied also on complex networks and, in theory, it can be run to infer networks of any dimension [Margolin et al., 2006];
- PC and ARACNE do not require *a priori* assumptions, such as unrealistic network models or data derived from perturbation experiments;
- PC and ARACNE use different strategies to find interactions between the variables.

The ARACNE algorithm (Section sec-ch2-NetwInfAlgo) was downloaded from `http://amdecbioinfo.cu-genome.org/html/caWorkBench/upload/aracne.zip`. Instead for the PC algorithm we used the Rpackage pcalg [Kalisch et al., 2010] that is publicly available (with open source code). The different strategies used by PC and ARACNE, to infer the interactions between the variables, is summarized in Figure 4.1.

4.2.1 Local Gene Network (LGN)

The choice to use *in vivo* data and a real LGN for the test implies the selection of an organism genetically well characterized, namely an organism for which the genes of the genome are known, gene expression data are available and some of its LGNs have been

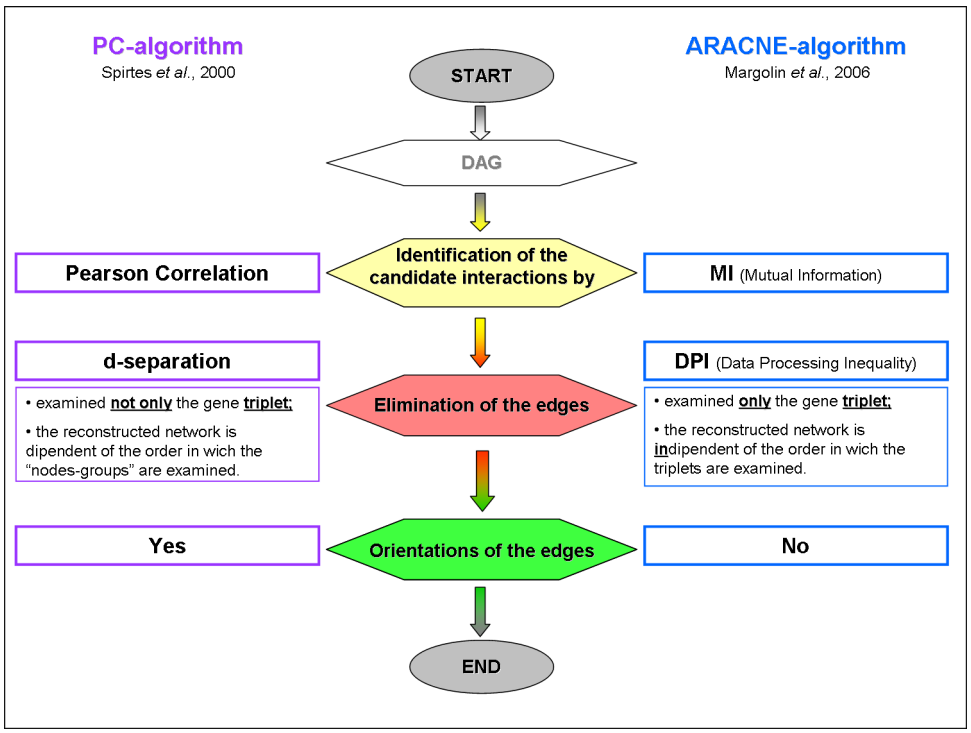


Figure 4.1: **Scheme of the different strategies used by the PC and ARACNE algorithms.**

validated by laboratory experiments. For these reasons, we have selected, as the organism *Arabidopsis thaliana*; the model for higher plants.

A known LGN of *Arabidopsis thaliana* is the Floral Organ Specification (FOS-GRN) which has been characterized and validated with specific mutants [Espinosa-Soto et al., 2004] [Sanchez-Corrales et al., 2010]. FOS-GRN comprises 15 genes and 54 causal relationships (Figure 4.2-c). This gene network is involved in the early differentiation of the inflorescence and of the floral organs and it integrates all previous knowledge on the ABC model [Coen et al., 1991]. The ABC model says explain how the so-called ABC-genes (AP1, AP2, AP3, PI and AG) as single activity or in combination cause the differentiation of the floral organs (sepals, petals, stamens and carpels). The ABC genes and other essential genes of this process (for example the SEP genes SEP1, SEP2 and SEP3) [Jack, 2001] are included in FOS-GRN (Figure 4.2-b). The ABC-genes are divided in three classes (A, B and C) with different activity. Genes of Class A are AP1 and AP2, genes of Class B are AP3 and PI and gene of Class 3 is AG. In particular the activity of genes of Class A specify the sepals; activity of genes of Class A and B specify petals and genes of Class B and C specify stamens (Figure 4.2-b). The ABC genes need also the SEP genes. All the three SEP genes are necessary for petals, sepals and carpels development, while SEP1 and SEP2 are enough for sepals development.

4.2.2 Gene Expression Data

As *in vivo* data we have chosen the observational gene expression data from *Arabidopsis thaliana* available at the Plant Expression Database (PLEXdb) [`http://www.plexdb.org`]. The final dataset contains 9 observational microarray experiments for a total of 393 hybridization with the *Arabidopsis thaliana* GeneChip®Arabidopsis ATH1 Genome Array (Table 4.8) (TIGR) which is based on the Affymatrix platform and contains 22,500 probe sets representing approximately 24,000 gene sequences. The advantage of selecting gene expression data from a single database (PLEXdb) was the fact of having all the data normalized with the same protocol.

4.2.3 Geneset Generation

In order to expand a LGN, PC-IM subdivides the whole gene list given in input (e.g. the genes of whole genome or those present on the microarry) in *tiles*, each one containing the genes belonging to the LGN and the other genes extracted randomly from the input gene list. This entire operation is repeated i times (i: number of iterations). This strategy allows:

- to use, in the step 2 of PC-IM, also infered algorithms that can analyse a limited

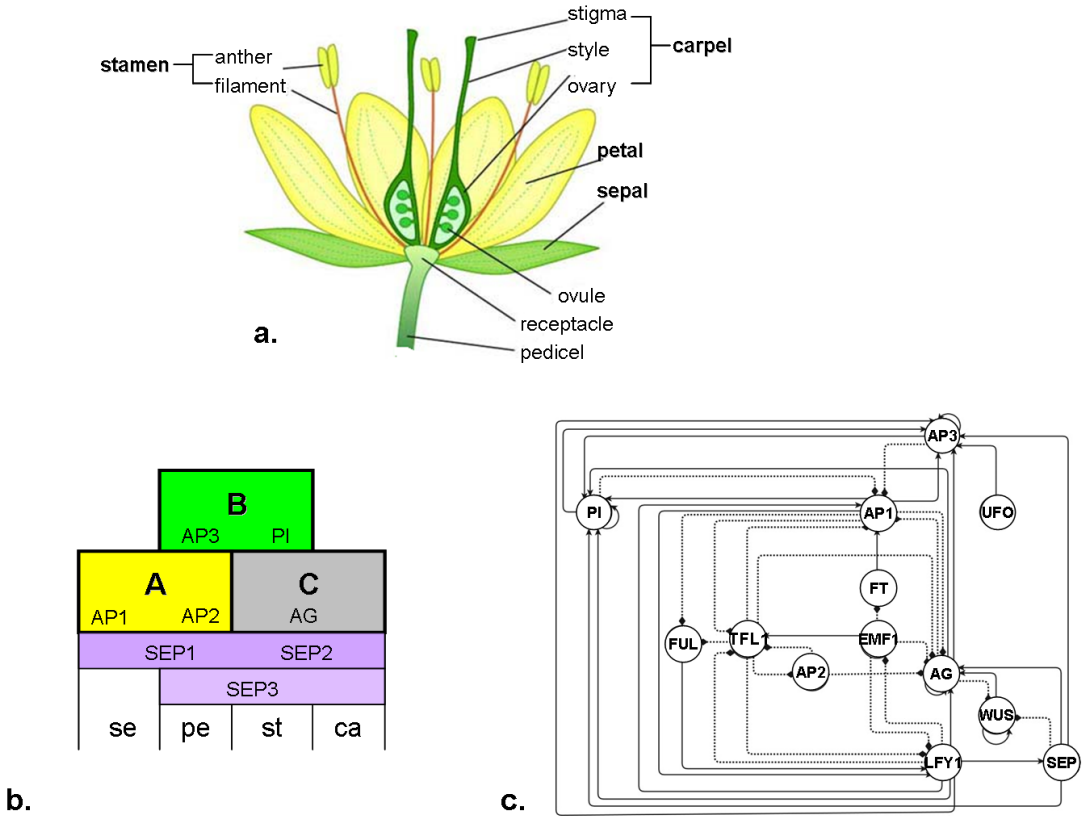


Figure 4.2: **Representation of the flower organs, ABC model and FOS-GRN of *Arabidopsis thaliana*.**

a. The flower organs. The words in bold are the organs specified by the ABC model. **b.** The ABC model (labeled boxes) and SEP genes (violet boxes). **c.** The FOS-GRN controls the early differentiation of inflorescence and floral organs in *Arabidopsis thaliana*. Positive and negative interactions are represented by continuous and discontinuous arrows, respectively [Sanchez-Corrales et al., 2010]. se: sepal, pe: petal, st: stamen, ca: carpel, AG: AGAMOUS, AGL8: AGAMOUS-LIKE 8, AP1: PETALA1, AP2: APETALA 2, AP3: APETALA 3, EMF1: EMBRYONIC FLOWER 1, FT: FLOWERING LOCUS T, LFY: LEAFY, PI: PISTILLATA, SEP: SEPALLATA, TFL1: TERMINAL FLOWER 1, UFO: UNUSUAL FLORAL ORGANS, WUS: WUSCHEL

Name	Type	Number
AT4	*Arabidopsis thaliana* gene expression during floral transition and early flower development	40
AT6	Loss of a callose synthase results in salicylic acid-dependent disease resistance	16
AT8	Expression analysis of Arabidopsis suspension cells during sucrose starvation	9
AT10	*Arabidopsis thaliana* gene expression after 6 days in shoot induction medium	30
AT12	Gene expression and carbohydrate metabolism through the diurnal cycle	22
AT13	Impact of Type III effectors on plant defense responses	27
AT17	Indole acetic acid treatment-dose response and time course	24
AT18	The mechanisms involved in the interplay between dormancy and secondary growth in Arabidopsis	36
AT40	Expression Atlas of Arabidopsis Development (AtGenExpress)	189

Table 4.8: **Description of the gene expression experiments from PLEXdb used to test the PC-IM.**

number of variables (e.g. the PC algorithm performans at best with 1000 variables [Wang et al., 2010]);

- to change the LGN surrounding;

- the parallelization of the algorithm which runs on different *tiles* in different iterations.

For the preliminary evaluation 2, five subsets of the ATH1 genes (called *tiles*) were created with different dimension: 50, 100, 200, 500 and 1000 genes. Each of this *tile* includes the 15 genes of the FOS-GRN and other additional genes randomly selected from the GeneChip®Arabidopsis ATH1 Genome Array. The number of these additional genes is different in the different *tiles*. In fact the tiles with 50, 100, 200, 500 and 1000 genes incorporate respectively 35, 85, 185, 485 and 985 random genes.

4.2.4 subLGNs Generation and Performances Evaluation

To evaluate the algorithms performances for the LGN expansion we would need a Gold Standard. In the PC-IM this problem is solved by choosing as LGN a well characterized gene network, dividing it into subLGNs of the same size and looking at the performance parameters (PPV, Se, Sp, ROC curve and PR curve) between nodes of the subLGN and other LGN genes not present in the subLGNs (for details: Chapter 3-step4 of the PC-IM).

The same procedure was used to test PC and ARACNE algorithms. In particular, 100 subLGNs of three different sizes (3, 5 and 10 genes) have been created. As in the step

4 of PC-IM, the subLGNs derive from the LGN and were obtained extracting randomly the genes of the FOS-GRN. For each subLGN, the genes of the LGN are divided in two categories: INTRA and EXTRA genes. For each size of subLGN we run 100 replicates.

The evaluation criteria are: Positive Predictive Value (PPV or Precision; Formula 3.4), Sensitivity (Formula 3.5), False Positive Rate (FPR; Formula 3.6), Number of total genes (Formula 4.1) and Delta (Formula 4.2):

$$\textit{Number of total genes} = (TP_{intra} + FP_{intra} + FN_{intra} + TN_{intra}) + (TP_{extra} + FP_{extra} + FN_{extra} + TN_{extra}) \quad (4.1)$$

$$Delta = (TP_{intra} + FP_{intra} + FN_{intra} + TN_{intra}) - (TP_{extra} + FP_{extra} + FN_{extra} + TN_{extra}) \quad (4.2)$$

4.2.5 Results

We have used subLGNs of 3 different size (3, 5 and 10 genes) and 5 different size od *tiles* (50, 100, 200, 500 and 1000 genes). There are 4 repetitions for each different size of the *tiles* and 100 replicates for each size of the subLGN.

Figure 4.3 shows the results of this preliminary evaluation 2 (Section 4.2). Figure 4.3-a shows the total number of genes found by the algorithms. With *tiles* of 50 genes, PC and ARACNE find a comparable number of genes, instead with *tiles* of other sizes ARACNE finds more nodes than PC. Figure 4.3-b represents the Delta criterion (Formula 4.2) to compare the two algorithms. The best performances are when the Delta value is positive or less negative, because this means that the algorithm finds more intra nodes than extra nodes. PC has Delta values close to zero for any *tile* size, while ARACNE is more variable and in general displays negative values. The difference in the Delta value is more evident with *tiles* of 500 genes. In this case, ARACNE has the most negative Delta value and the biggest difference with PC. Figure 4.3-c and Figure 4.3-d show PPV (Formula 3.4) and Se (Formula 3.5) respectively. The PPV of PC is greater of ARACNE, except in the case of *tiles* with 1000 genes (Figure 4.3-c). The Se of ARACNE is greater then that of PC (Figure 4.3-d), except in the following cases:

- with subLGN with 10 genes and *tiles* with 200 and 1000 genes, PC and ARACNE have similar Se values;
- with subLGN with 5 genes and *tiles* with 500 genes , PC and ARACNE have similar Se values;
- with 10 genes in the subLGN and 500 genes into the *tiles* PC has greater Se with respect to ARACNE.

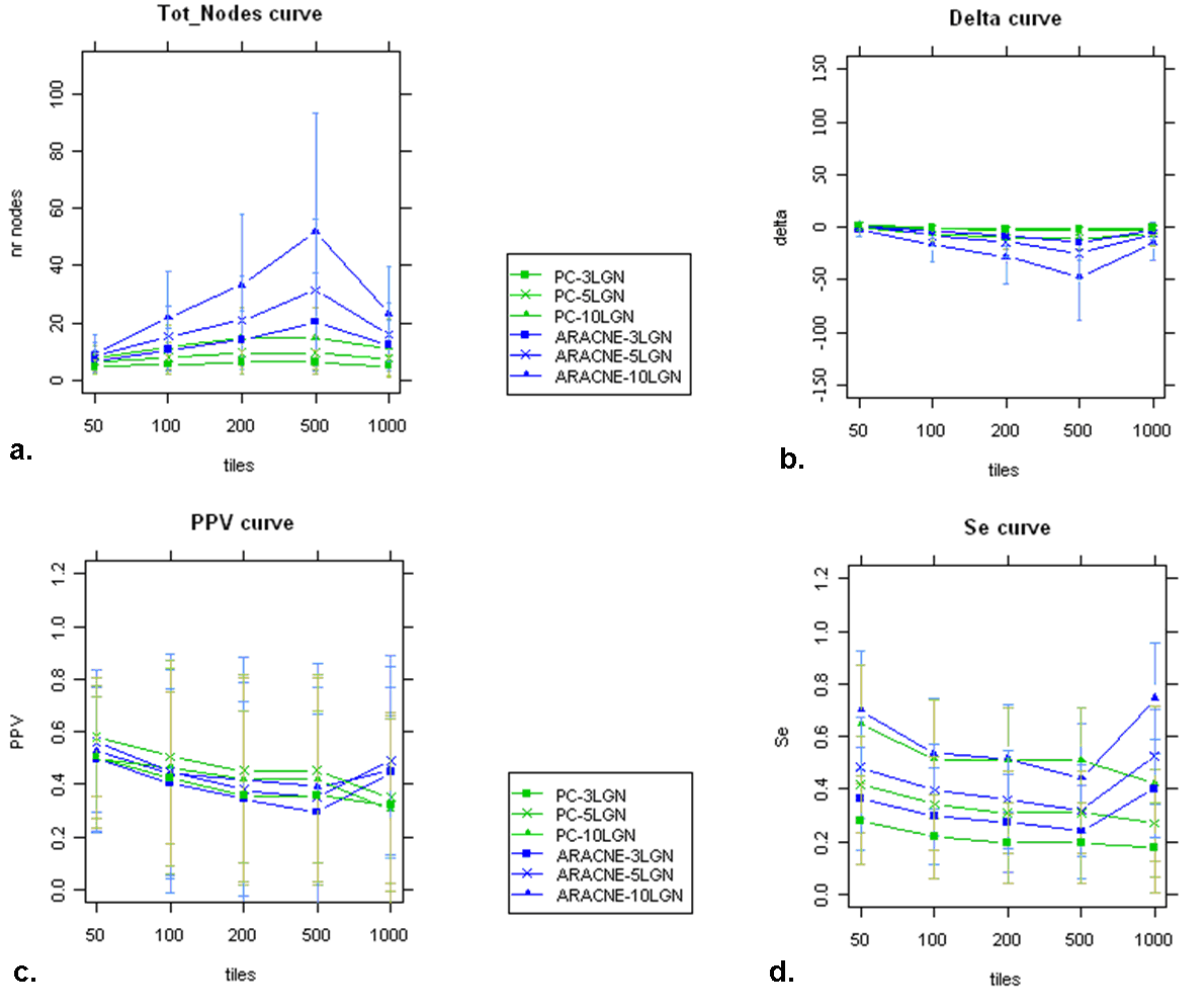


Figure 4.3: **Results of the preliminary evaluation 2.**
In green is the PC algorithm and in blue is ARACNE. The different lines are the different size of the LGN (3, 5 and 10 genes, represented by cube, asterisk and triangle respectively). **a.** Tot-Nodes curve: total number of genes found by the algorithms **b.** Delta curve: the Delta criterion **c** PPV curve **d** Se curve.

4.2.6 Discussion of the preliminary evaluation 2

The expansion nodes for a certain LGN, found by an *in silico* approach, need to be followed by an *in vivo* validation. This step which consists in complex laboratory experiments (for example yeast two-hybrid experiments for the validation of paired physical interaction, or the generation of genetic mutations) due to the time and costs required can be afforded only for a limited number of genes. Moreover both PC and ARACNE do not have 100 % precision and sensitivity values (see results of preliminary evaluation 1, Figure 4.3-PPV curve and Figure 4.3-Se curve). In case the performances in terms of precision (PPV) between different algorithms are comparable (e.g. Figure 4.3-PPV curve and Figure 4.3-Se curve with some size of *tiles*) then is preferable to choose the algorithm that finds the smaller number of positive genes (Figure 4.3-Tot_Nodes curve) and has a major value of Delta. The combination of these two parameters (number of total genes and Delta) improves the probability to choose true expansion genes, limiting the insuccess rate validation experiments.

4.3 Evaluation of PC-IM

PC-IM performances have been evaluated using the same *in vivo* gene expression data and LGN of Section **??**, namely 393 hybridization experiments from *plexdb* database and FOS-GRN of *Arabidopsis thaliana* LGN. The aim of this evaluation was to understand whether the performances of the method are depending on input gene expression data on type of LGN or on the other parameters (i.e. number of iteration (i) and *tile* size (t)). For these reasons the following evaluations will consider the:

- Effect of the *tile* size t;
- Effect of the iteration number i;
- Effect of the frequency (value determined in step 5 of PC-IM);
- Effect of the gene expression type;
- Effect of the LGN (random LGN vs real LGN).

Finally we compared the performances of PC-IM with those of GENIES [Kotera et al., 2012].

4.3.1 Effect of the *tile* size t

In step 1 of PC-IM the input gene list is divided into *tiles* of a size t that is specified by the user. In this section we investigate the changes in PC-IM output with different values

of t and we identify the t value that gives the best performances.

This analysis is made on five different *tile* sizes: 50, 100, 200, 500 and 1000 genes. The number of iteration i is 100 and the dimension of the subLGN d is 10. Figure 4.4 shows the results of this test and it indicates that the lowest d_{min} in the PR curve found for t = 50 genes (red line) and t = 100 genes (magenta line) (Figure 4.4-PR curve and Table 4.9). Instead the biggest values of AUC are associated with *tile* sizes of 200, 500 and 1000 genes (Figure 4.4-ROC curve and Table 4.9). Nevertheless the differences between the biggest and smallest values of AUC of the ROC curves and d_{min} of the PR curves are in the order of the hundredths (Table 4.9). This means that the size of the *tile* does not affect the methods performances and that the *tiles* can have size ranging between 200-1000 genes.

	50 genes	**100 genes**	**200 genes**	**500 genes**	**1000 genes**
AUC	0.662	0.663	0.710	0.706	0.700
d_{min}	0.481	0.496	0.547	0.560	0.562

Table 4.9: **Value of AUC and d_{min} with different *tile* size.**
The size of the *tile* (number of genes) is indicated in the first row.

In the next steps of PC-IM evaluation we used a fixed *tile* size of 1000. This choice derives from the above considerations of the AUC and d_{min} values and from the results (Figure 4.3-PPV curve) with different subLGNs sizes which showed that this *tile* size is not affected by changes in the dimension of subLGN.

4.3.2 Effect of the number of iterations *i*

A parameter that have to be set is the iterations number i that is the number of times the procedure of the *tiles* generation from the input gene-list, is repeated. This parameter permits:

- to increase the number of combinations between intra and extra genes, into the *tiles*;

- to parallelize step 1 to step 4 of the algorithm.

To study the effect of the number of iterations i on PC-IM results we used different i values (1, 5, 10, 15, 20, 25, 50, 75 and 100), d = 10 (where d is the size of the subLGN) and t = 1000. At the end of this analysis we will choose the i value that is the best compromise between biggest AUC and smallest $_{dmin}$.

In Figure 4.5 the ROC and PR curves for this test are presented. The results show that at similar values of percentage frequency, 1-Sp (Figure 4.5-ROC curve) and PPV (Figure 4.5-PR curve) are comparable and not affected by the number of iterations. Instead the Se value increases with the increase of the number of iterations. This means

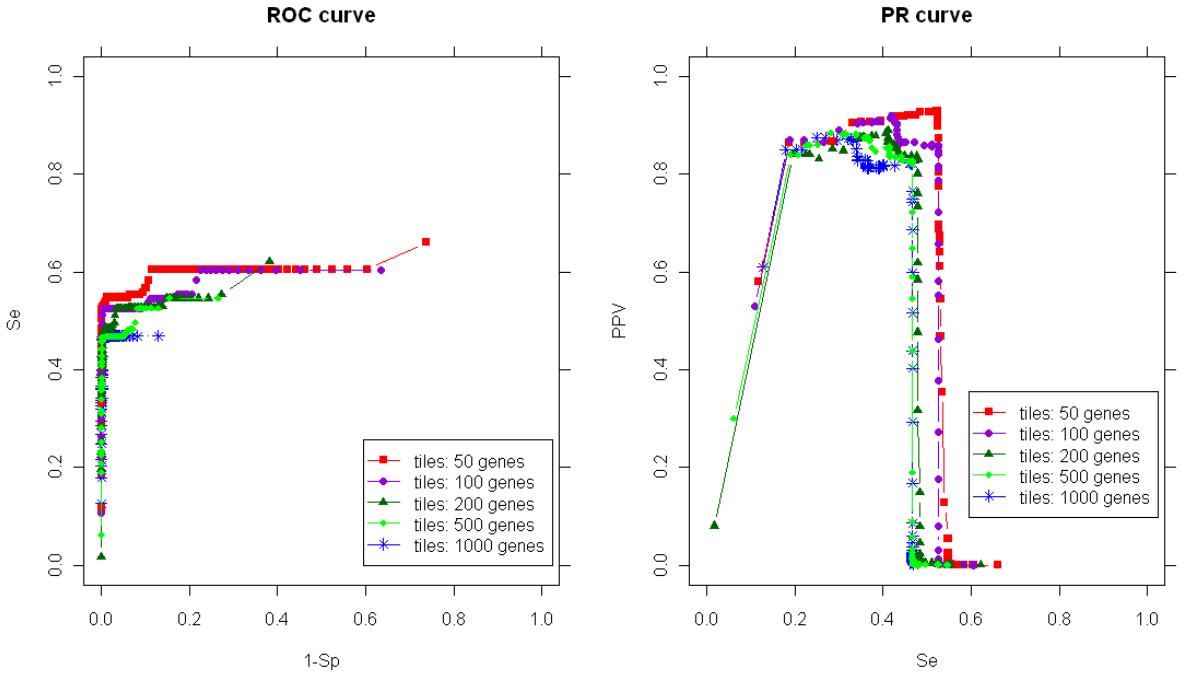


Figure 4.4: **ROC and PR curve of the *tile* size *t* effect.**
Each colour is a different size of *tile* t and each point is a different percentage value of frequency % (frequency calculates in the step 5 of the PC-IM).

that the number of iterations has an higher effect on the number of false negatives with respect to the number of true positives and increasing the i value decreases the number of false negatives, but it does not change significantly the number of true positives, (PPV is comparable with different i). The reason for the decrease of false negatives is that with high i, the number of the total *tiles* to be tested increases. With total *tiles* we mean the number of the *tiles* for an iteration multiplied by the number of total iterations. The greater the number of the total *tiles*, the higher will be the number of different genes combinations (intra and extra) as well as the selectivity of the method. This permits PC-IM to reduce the number of false negatives.

Number of i	1	5	10	15	20	25	50	75	100
AUC	0.730	0.747	0.727	0.683	0.683	0.719	0.713	0.683	0.700
d_{min}	0.834	0.676	0.738	0.723	0.724	0.611	0.573	0.533	0.562

Table 4.10: **Values of AUC and d_{min} with different iteration number**
The size of the *tiles* (number of genes) is indicated in the first row.

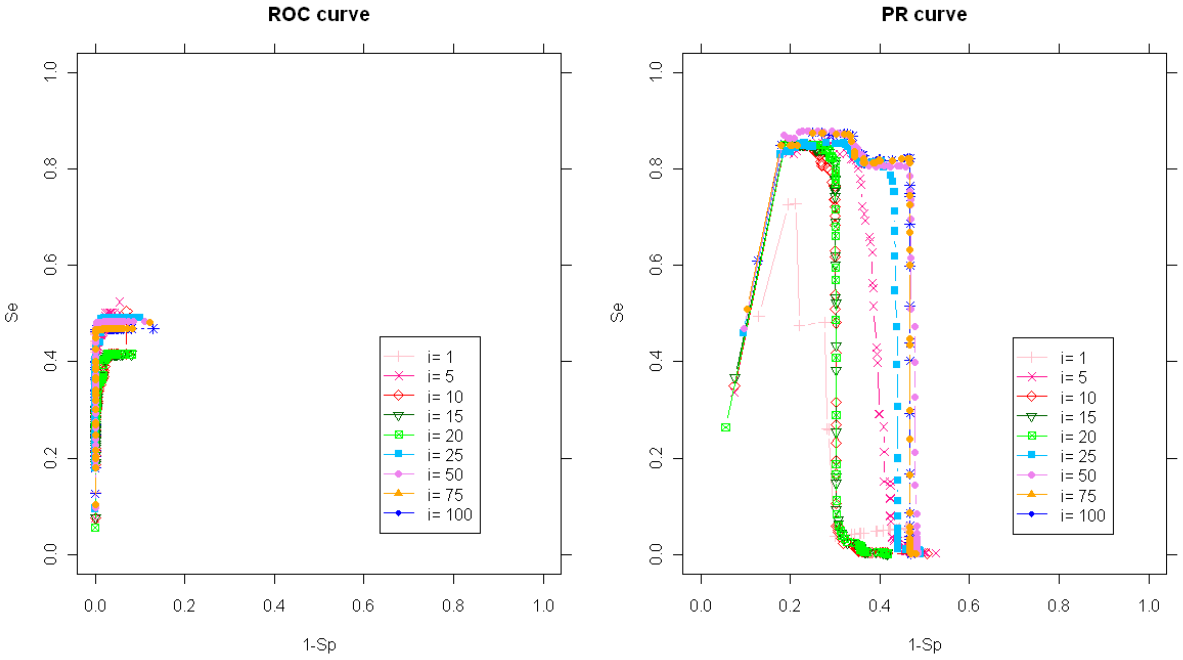


Figure 4.5: **ROC and PR curveof iteration number i effect.**
Each colour is a different number of iteration and each point corresponds to a different percentage value of frequency % (frequency calculates in the step 5 of the PC-IM).

In this test $i = 100$ is the iteration value that has the best compromise between the biggest value of AUC and smallest value of d_{min} (Table 4.10) and for this reason will be our choice for the future expansion of the LGN.

4.3.3 Effect of the type of gene expression data

In the preliminary results 1 (Section 4.1) it was shown that the type of gene expression data has a great influence on the performances of the algorithms. This happened both when comparing *in silico* data *versus in vivo* data and when comparing two different type of *in vivo* data. In this section we test if this effect is observed also with PC-IM. For this reason PC-IM is used on different combinations of *in vivo* expression data. These combinations are called SubSets and are generated starting from the 393 gene expression data described in Section 4.2.2. These SubSets are different for the number of hybridization experiments and for the presence or not of gene expression data related to the LGN to be expanded. We analysed three SubSets (SubSet A, SubSet B and SubSet C) with the following composition:

- SubSet A is formed by the AT4, AT6, AT8, AT10, AT12, AT13, AT7, AT18 and

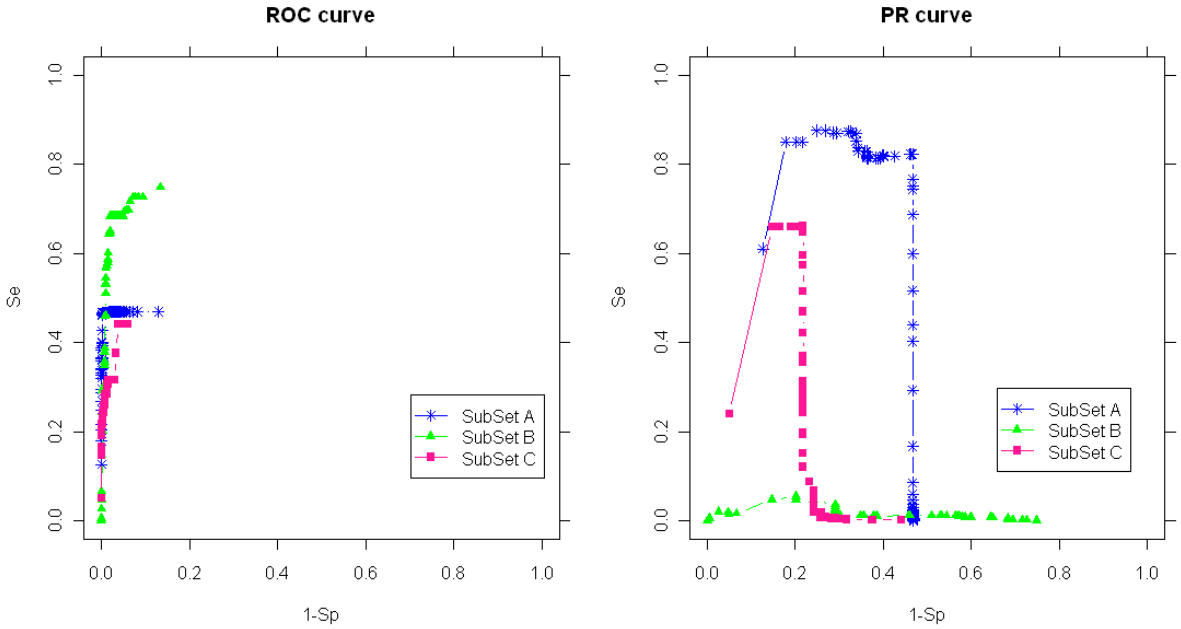


Figure 4.6: **ROC and PR curves for the dependence on different SubSets A, B and C.** Each colour is a different set of gene expression data (SubSet) and each points is a different percentage value of the frequency (frequency calculates in the step 5 of PC-IM).

AT40 hybridization experiments;

- SubSet B is formed by the AT6, AT8, AT10, AT12, AT13, AT7, AT18 and AT40 hybridization experiments;

- SubSet C is formed by the AT4 hybridization experiments.

Table 4.11 reports the performances of PC-IM with the different SubSets. The evaluated parameters are the cut-off frequency, AUC and d_{min}. The cut-off frequency indicates the value of frequency calculated and used by PC-IM to expand the LGN. The AUC and d_{min} reported in the table correspond to the selected cut-off frequency. In this test the LGN is the FOS-GRN, $d = 10$, $t = 1000$ and $i = 100$.

The results showed in Table 4.11 and in Figure 4.6 clearly indicate that PC-IM gives different outputs with different gene expression data.

With 353 gene expression datasets not including data related to the flowering process, we have obtained the best value of sensitivity (Figure 4.6-ROC curve: green line) but the worse value of PPV (Figure 4.6-PR curve: green line). This suggests that with expression data unrelated to the studied LGN, PC-IM finds a high number of expansion genes and consequently less false negatives, but also more false positives.

	SubSet A	**SubSet B**	**SubSet C**
Type of Hybridization Experiments	Flowering + noFlowering	noFlowering	Flowering
Total number of Hybridization Experiments	393	353	40
cut-off frequency	61.0	1.0	71.5
AUC	0.533	0.285	0.782
d_{min}	0.562	1.030	0.852

Table 4.11: **PC-IM performances with different gene expression data (SubSets A, B and C).**
The term "Flowering "indicates gene expression data related to the flowering process, instead "noFlowering"means unrelated with flowering.

In the case of 393 hybridization experiments (flowering related and not related expression data) the value of sensitivity (Figure 4.6-ROC curve: blue line) is comparable with the one of SubSet C (40 hybridization experiments specific of the flowering process) (Figure 4.6-PR curve: magenta line), but PPV is better than that observed with the other two SubSets (Figure 4.6-PR curve). These results show how the presence of the non-specific biological experiments helps PC-IM to reduce the number of false positives.

4.3.4 Effect of the LGN (Real LGN vs Random LGN)

In this step we want to test the robustness of PC-IM comparing its performances in the inference of two different LGNs and the relative subLGNs. The first LGN we have chosen is the FOS-GRN, namely a real biological LGN. The second LGN is a random LGN (Random LGN), obtained by randomly selecting the LGN genes from the genes of the input dataset. Real LGN and Random LGN have the same size (number of genes) and the same number of relationships. In the Random LGN, also the relationships between the genes of the LGN are obtained randomly. In particular, in this case, we have originated three different combinations of 54 relationships among genes, in order to have three repetitions.

The input gene expression dataset is formed by the 393 hybridizations of the ATH1 GeneChip, number of iterations $i = 100$ and *tile* size $t = 1000$. The results are show in Table 4.12. The results clearly demostrate that PPV, Se and cut-off frequency obtained with FOS-GRN are much greater than those obtained with the Random LGN. In addition d_{min} observed for FOS-GRN is lower with respect to the d_{min} observed for the Random LGN.

These values indicate that PC-IM is influenced by the nature of LGN. In fact when we use the Random LGN, PC-IM finds expansion genes with lower precision and sensitivity.

	PPV max	s.d.	**Se max**	s.d	d_{min} **(PR curve)**	**cut-off frequency**
FOS-GRN	0.82230	-	0.46700	-	0.56200	62
Random LGN	0.00016	± 8.37e-06	0.09133	± 0.004619	1.35100	4

Table 4.12: **PC-IM performances in expanding FOS-GRN or a Random LGN.** (s.d. is standard deviation).

4.3.5 Effect of the frequency value

The choice of the cut-off frequency (step 4 of PC-IM) determines the final LGN expansion gene list (step 5 of PC-IM). The selection of the cut-off frequency is done by identification of the frequency value that maximizes AUC (from ROC curve) and minimizes d_{min} (from PR curve). When it is not possible, to reach the maximun and the minimun of these two parameters then the cut-off frequency is the best compromise to optimize them.

To test the quality of the output selected from the cut-off frequency, a validation of the final expansion gene list was necessary. This validation was based on a bibliographic search that compared the genes provided by PC-IM in the expansion list with the information present in the literature. According to the bibliographic search we divided the genes, of the expansion list, in four classes:

- Class 1: genes related to the LGN;
- Class 2: genes not directly related with the LGN, but related with genes into the Class 1;
- Class 3: genes unrelated with the LGN;
- Class 4: genes not supported by references.

In Class 3 is important to clarify that there are genes involved in metabolic processes different to the metabolic process in which the LGN is taking part according to the state of art knowledge. This does not exclude completely that those genes are related with the genes of the LGN. In fact, we can say only that in Class 3 there are genes that do not have, in the literature, evidence to be related with the LGN. For Class 4, instead, we can not make any consideration, because in literature there is no information that help us to understand if PC-IM expansion genes are correct.

For this evaluation we run PC-IM on the FOS-GRN LGN using the following parameters: $t = 1000$, $i = 100$, $d = 10$. The results are shown in Figure 4.7. The Se and PV curves (Figure 4.7-PPV and Se curve) depict the variation of PPV and Se with different values of frequency (0-100). The black line indicates the frequency selected, by PC-IM,

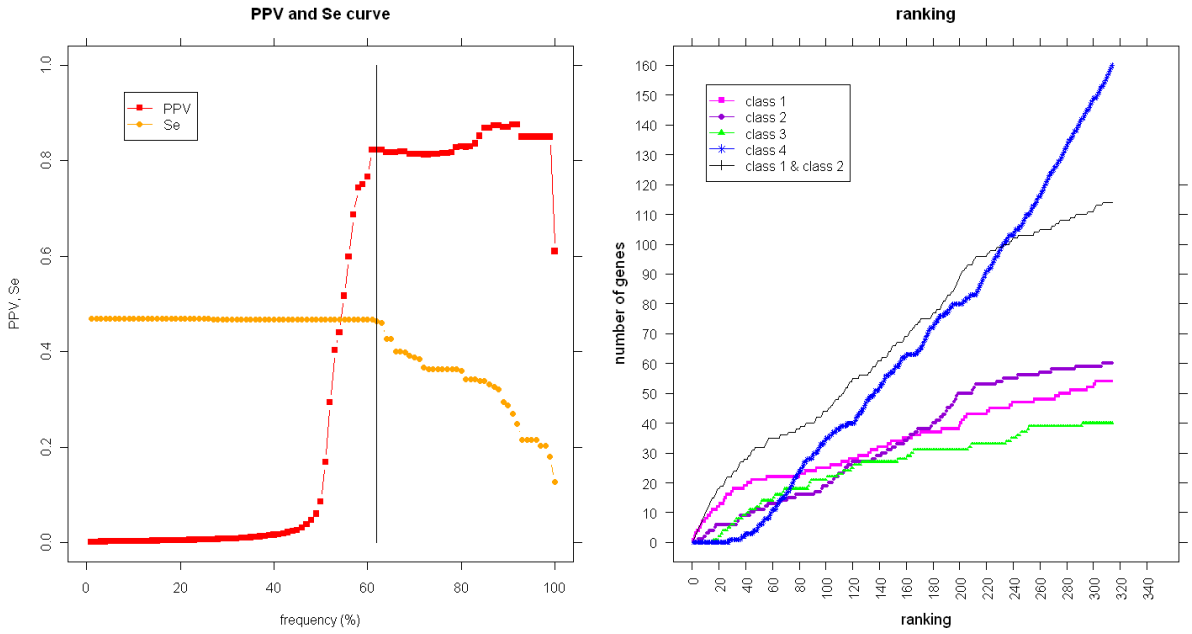


Figure 4.7: **PPV-Se curve and ranking of the FOS-GRN expansion.**

to have the final expansion gene list. Figure 4.7-ranking shows the distribution of the expansion gene list (314 genes) into the four classes (Class 1, Class 2, Class 3 and Class 4). The expansion list is the output of PC-IM and it is obtained using 62% as cut-off frequency. This value derives from the intrinsic evaluation of the PC-IM (black line in Figure 4.7- PPV and Se curve).

	Class1	**Class 2**	**Class 3**	**Class 4**
number of genes	54	60	41	159

Table 4.13: **Distribution of the expansion FOS-GRN genes into four classes.** (Class 1 = genes related with flowering; Class 2 = genes not directly related with flowering, but with Class 1; Class 3 = genes Unrelated with flowering; Class 4 = genes without references).

The cut-off frequency obtained by PC-IM is 62% and at this frequency the total number of expansion genes is 314. The distribution of these 314 genes into the four Classes is showed in Table 4.13.

At the highest values of frequency the total number of genes in the PC-IM output is very low (the sensitivity estimated from PC-IM is around 10-20%) (Figure 4.7-PPV and Se curve). Moving towards lower frequency values will increase the sensitivity, but will not affect the PPV values until a frequency value of 62% is reached. At lower frequency

indeed, PPV decreases dramatically, while Se reaches a plateau. PC-IM will then select the 62% frequency as cut-off frequency.

The top part of the ranking (higher frequency values) is mainly populated by genes of Class 1 followed by those belonging to Class 2 and 3, until genes of Class 4 are very few (Figure 4.7.ranking). This behaviour is a clear indication that the genes at the top of the expansion list are very likely to bestrongly related to the LGN.

In the lower part of the ranking the situation changes. The genes in Class 4 start to be included at higher rate than the genes of the other Classes, in the output list.

In this particular case, FOS-GRN, the turning point is at about position 40 of the ranking.

It is worth to mention, however, that the proportion of "true "genes (Class 1 + Class 2) remain higher than that of "false"genes until position 220 of the ranking.

4.3.6 Comparison of PC-IM *versus* GENIES

As a last step of the evaluation procedure we decided to compare PC-IM with the competitor method GENIES, a recently published method for LGN expansion.

The analysis compares PPV, Se, Sp, ROC and PR curves. To evaluate the algorithms' performances, the LGN was divided in 100 subLGNs. Each subLGN was obtained randomly choosing genes from the LGN. The genes in the subLGN are divided in two classes: INTRA and EXTRA. The INTRA genes are those genes that belong to the LGN, instead the EXTRA genes are those present in the input data, but are not included in the LGN. Once obtained the Gold Standards the evaluation criteria are obtained using Equations 3.4, 3.5 and 3.5 to calculate PPV, Se and Sp respectively.

For this comparison we have chosen three different LGNs, called LGN 1, LGN 2 and LGN 3. These LGNs differ for their size and for the organism they belong. The description of the three LGNs is reported below and summarized in Table 4.14.

	LGN 1	**LGN 2**	**LGN 3**
Organism	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*	*Saccharomyces cerevisiae*
Size of the LGN	15	133	14
Size of the subLGN	10	86	9
Number of subLGNs	100	100	100
Expression array [genes; experiments]	[22810; 393]	[544; 157]	[3370; 397]

Table 4.14: **Description of the three different LGNs used to compare the performances of PC-IM and GENIES.**

LGN 1

The network is the FOS-GRN of the *Arabidopsis thaliana* (see section 4.2.1). The FOS-

GRN comprises 15 genes related with mutation experiments carried out in *Arabidopsis thaliana* [Espinosa-Soto et al., 2004] [Sanchez-Corrales et al., 2010]. The expression data derives from the plexdb database (`www.plexdb.org`) and consists of 393 hybridization experiments (see section 4.2.2). The parameters used to run PC-IM were *tile* size of 1000 genes and number of iterations equal to 100. For the comparison of the methods the size of the subLGNs was set to 10.

LGN 2

LGN 2 contains 133 genes of *Saccharomyces cerevisiae* and the expression data are formed by 157 experiments and 544 genes. LGN and gene expression data are available on the GENIES website (`http://www.genome.jp/tools/genies/help.html`) and are those used as example to explain the method.

In this case the parameters, to test PC-IM, were $t = 200$ and $i = 100$. Since the number of genes in the expression dataset are 544, a *tile* size of 200 genes permits to have a good number of different genes surrounding the LGN. This is important, because we have observed previously that PC-IM output is affected by the gene composition of the *tiles*. This is due to the use of the *d-separation* criterion (Chapter 2). The *d-separation* criterion finds the v-structures between the variables, then changing the combination of nodes makes possible to find new v-structures.

LGN 3

LGN 3 is formed by genes involved in the glycolysis pathway of yeast (*Saccharmoyces cerevisiae*). These 14 genes listed in Table4.15 have been taken from the website Saccahromyces Genoma Database (SGD) (`http://pathway.yeastgenome.org/YEAST/NEW-IMAGE?type=PATHWAY&object=GLYCOLYSIS&detail-level\=2&detail-level=3&detail-level=2`) and the pathway is showed in Figure 4.8. The expression data are 397, derived from the SGD database (`http://www.yeastgenome.org/download-data/expression`) and are listed in Table 4.16. The choice of the type of expression data was done randomly, except for GSE7820_set0_ family, GSE7820_set1_ family, GSE7820_setA_ family, GSE8898_setA_ family and GSE9232_setA_ family that are related to the glycolysis pathway. The number of expression data and the size of the LGN are comparable with the number of the gene expression data (393) and LGN size (15 genes of the FOS-GRN) of FOS-GRN expansion. The reason for this choice was to verify whether PC-IM givs comparable performances to those obtained in the expansion of FOS-GRN, when provided with similar input data (size of LGN and number of expression data). For PC-IM we have chosen to use $i = 100$ and two different sizes of *tile*: 200 and 500.

GENIES can be used on the website `http://www.genome.jp./tools/genies/`. Requirements are partial knowledge of the metabolic network (LGN or input data set) and any "profile" of genes (or proteins) (e.g. gene expression profile, protein subcellular local-

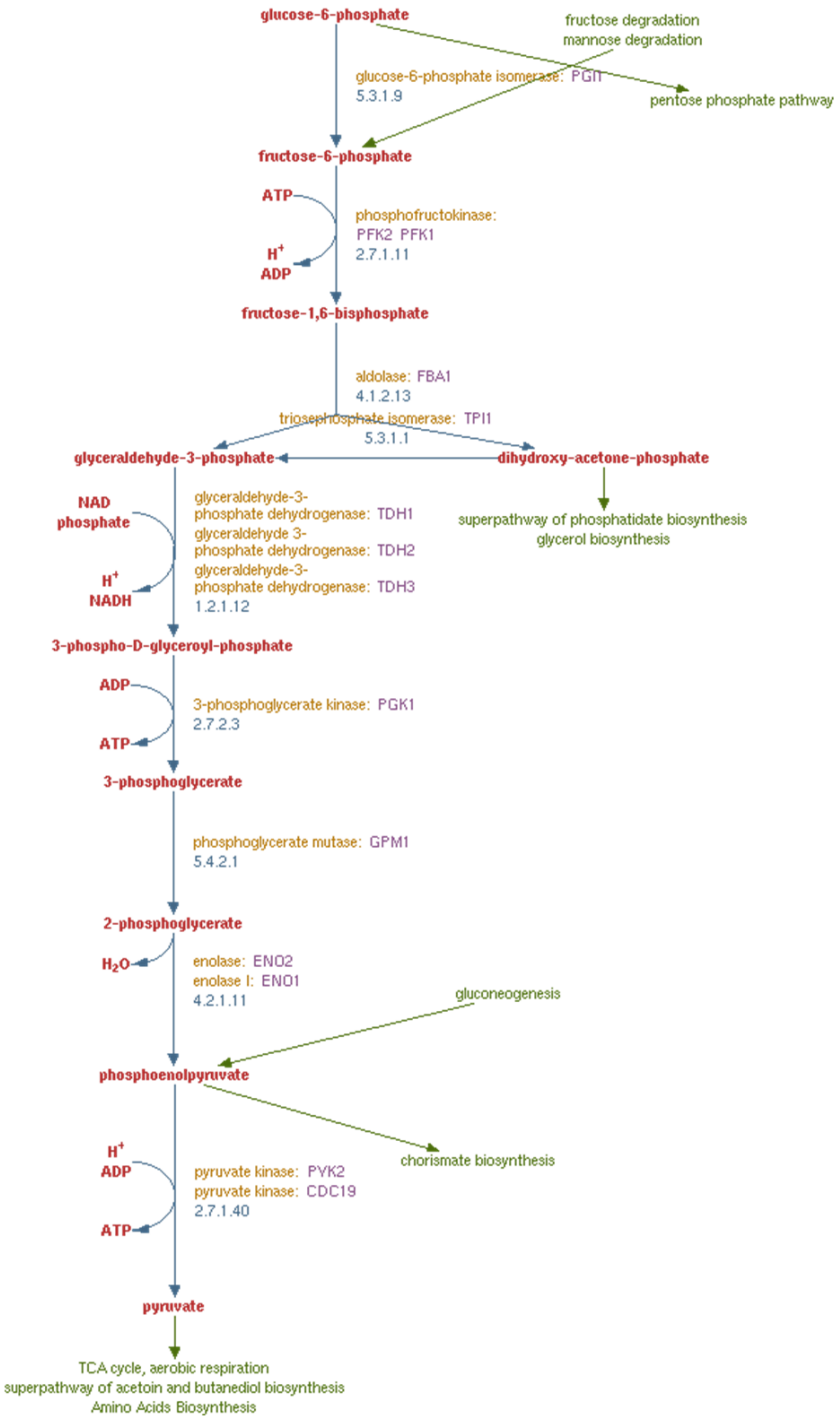


Figure 4.8: **The glycolysis pathway in *Saccharomyces cerevisiae*** (taken from SGD (Saccharomyces Genome Database) `http://www.yeastgenome.org/`).

Standard Name	Systematic Name
PGI1	YBR196C
PFK2	YMR205C
PFK1	YGR240C
FBA1	YKL060C
TPI1	YDR050C
TDH1	YJL052W
TDH2	YJR009C
TDH3	YGR192C
PGK1	YCR012W
GPM1	YKL152C
ENO2	YHR174W
ENO1	YGR254W
PYK2	YOR347C
CDC19	YAL038W

Table 4.15: **List of genes involved in the glycosilic pathway.**

ization profiles and phylogenetic profiles). In GENIES, each genes or protein input data set is transformed into the kernel similarity matrix by a kernel function [Kotera et al., 2012] and the user can select the kernel matrix and the algorithm to be used for LGN expansion. For this reason, we have compared PC-IM with different combinations of kernel matrix and algorithm. These combinations are listed in Table 4.17.

Figure 4.9 shows that GENIES_11 (GENIES with exponential kernel and penalized kernel matrix regression) GENIES_12 (exponential kernel matrix and em-algorithm) and GENIES_3 (exponential kernel matrix combined with kernel canonical correlation analysis) have had apparently better performances respect to PC-IM. The values of d_{min} of the GENIES_11, GENIES_12 and GENIES_3 are all 0.249, instead the value of d_{min} of PC-IM is 0.562. However in the gene expansion list (not selected with cut-off) of these three combinations of GENIES, there are no genes, while in the gene expansion list of PC-IM, before the selection by cut-off frequency, there are 4086 genes.

The minimum value of d_{min} in the expansion of LGN2 is 0.417. This value is obtained with GENIES algorithm and in particular with the combinations of kernel matrix and algorithms present in GENIES_3, GENIES_11 and GENIES_12 (Figure 4.10). The d_{min} value of PC-IM is 0.527. The three combinations of GENIES have in the final expansion list (before the selection with cut-off) only 1 gene, instead PC-IM gives 422 genes in the expansion list (without cut-off frequency selection).

In the case of LGN3 expansion the minimum value of d_{min} is 0.463 and this is obtained with PC-IM and $t = 500$. The GENIES combinations with minimum value of d_{min} (0.543) are GENIES_3, GENIES_11 and GENIES_12 (Figure 4.11). In this case PC-IM (t=500)

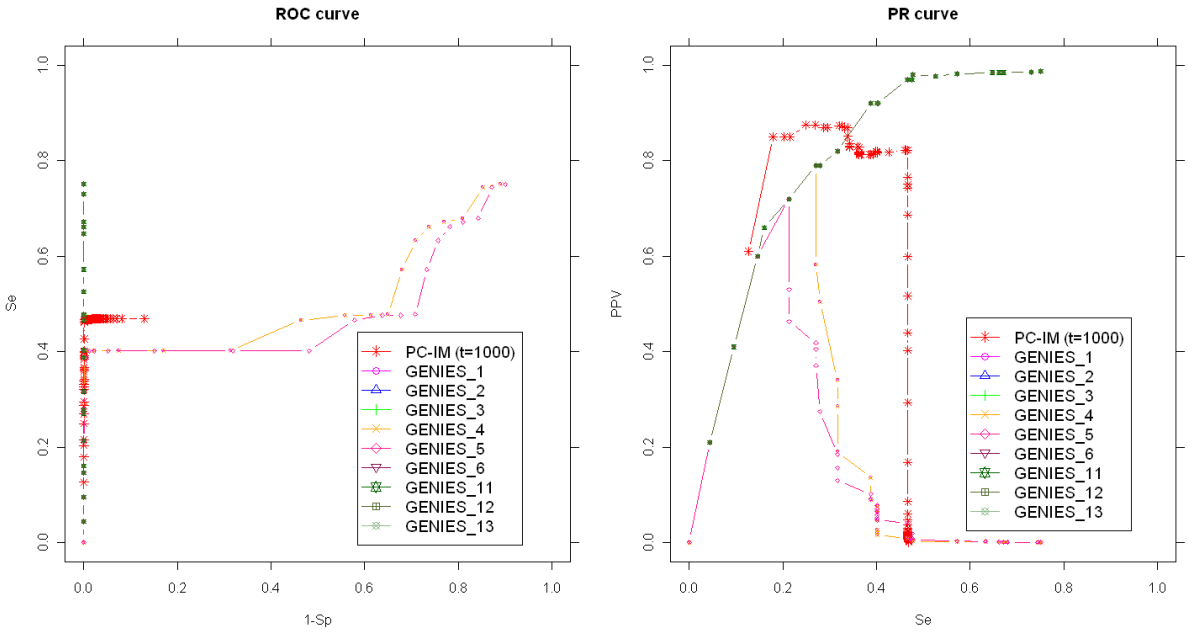


Figure 4.9: **ROC curve and PR curve of LGN1.**
Comparison of PC-IM with different version of GENIES see Table 4.17.

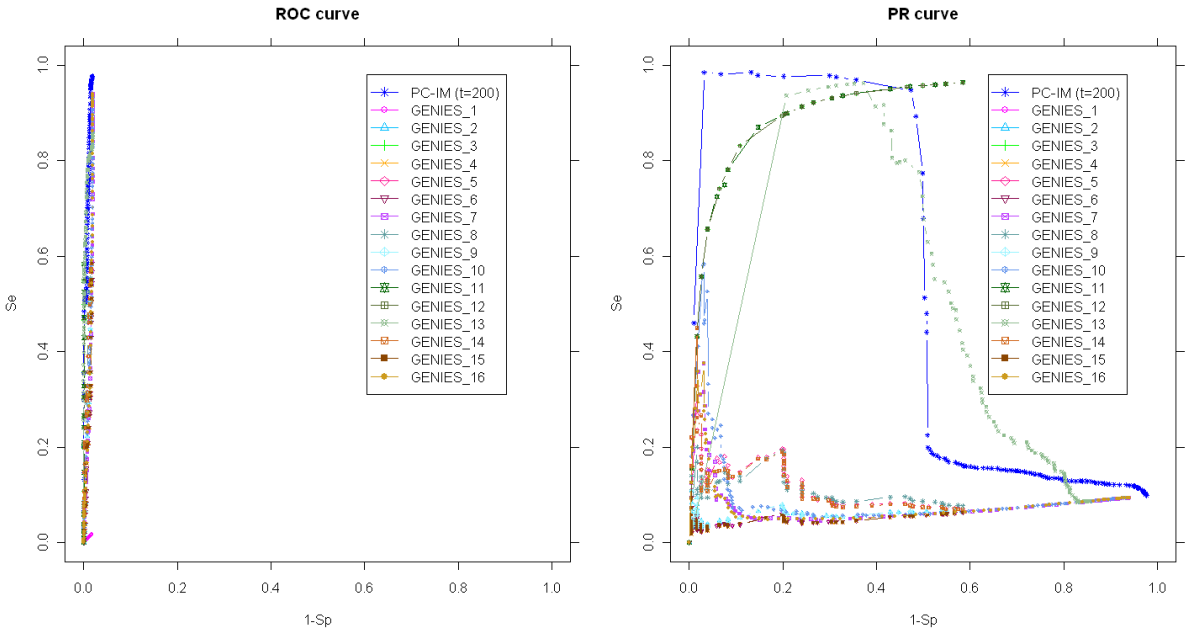


Figure 4.10: **ROC curve and PR curve of LGN 2.**
Comparison of PC-IM with different version of GENIES see Table 4.17.

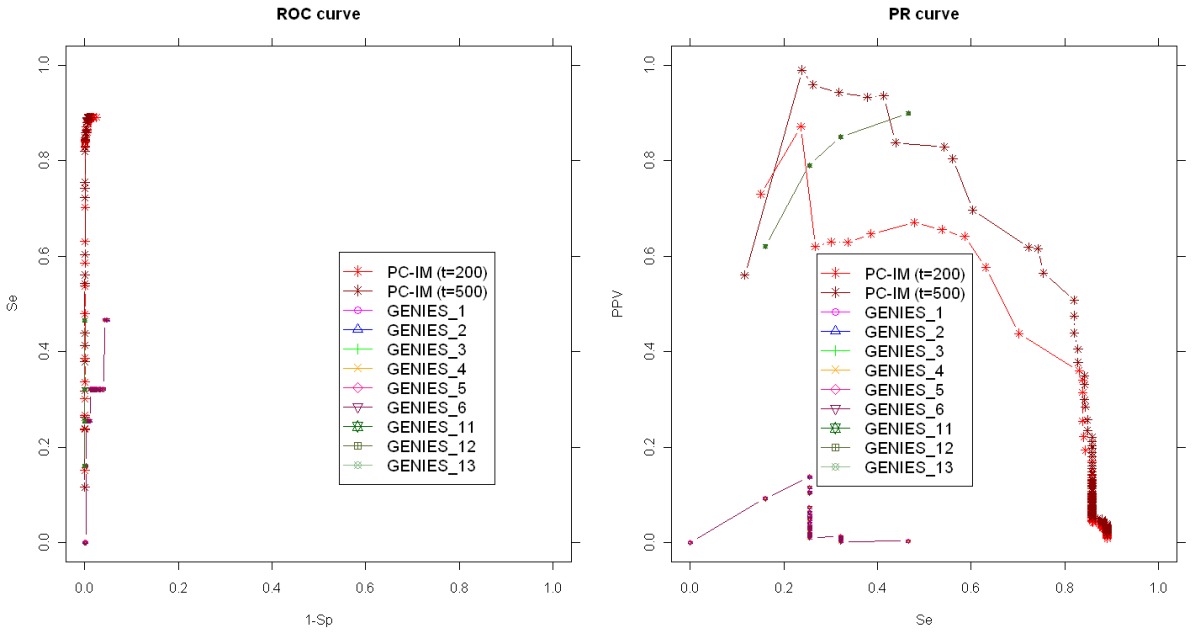


Figure 4.11: **ROC curve and PR curve of LGN 3.**
Comparison of PC-IM with different version of GENIES see Table 4.17.

produces a gene expansion list (without cut-off frequency selection) of 449 extra genes, instead in the expansion lists of GENIES_3, GENIES_11 and GENIES_12 there are 0 extra genes.

In conclusion PC-IM has performed much better than GENIES in finding genes outside the LGN.

4.3.7 Conclusion of the PC-IM Evaluation

PC-IM is a method to expand GRNs. Its evaluation comprised the analysis of the effect of the *tile* size (Section 4.3.1), the number of iterations (Section 4.3.2), the type and the number of the gene expression data (Section 4.3.3) and the LGN (Section 4.3.4). In addition an evaluation of the output data has been done to investigate the effect of the frequency (Section 4.3.5). This is an intrinsic parameter generated by PC-IM and used to output the final gene expansion list.

The subdivision of the input geneset in *tiles* permits to analyse very large genesets such as the whole genome of a species and to test different genes around the LGN's genes. The evaluation of the size of the *tiles* shows that 100 and 1000 are the number of genes that permits to obtain the best performances. This confirms the findings on the LPC (Low PC algorithm) reported by Wang et al. [2010]. In the range of *tile* size 100-1000 we

have chosen $t = 1000$ for two reasons. First, t =1000 gives the best compromise between ROC and d_{min} and second, with this value of *tile* size is possible to expand LGNs greater than 15 genes. In fact, in this case ($t = 1000$) we have more external genes with respect to the LGN.

The iterations have the aim to repeat the expansion on the whole geneset more times and than to select only the expansion genes that have been outputted in many iterations. This means that the iteration number is important to estimate of the cut-off frequency. We have found that a number of iterations equal or greater than 50 gives the best performances in terms of PPV and Se.

Gene expression data is another important parameter to be considered. Gene expression data can change for the their number (e.g. number of the hybridization experiments in a microarray) and for their type (they can be related or not with the LGN). In the choice of the type of gene expression data, there are three possibilities: all experiments are specific for the LGN (e.g. the hybridization experiments are related to the flower and the LGN is the FOS-GRN), the experiments regard the biological topic of the LGN or the experiments are a mix of two previous situations. These three categories using the gene observational gene expression data present in the plexdb database [`www.plexdb.org`]. This test has underlined how the best performances were detained by PC-IM when using a high number of experiments that were a mix of specific and not specific experiments.

To estimate the robustness of the methodology two different LGNs have been tested: a Real LGN (FOS-GRN) and a Random LGN. The Random LGN was obtained randomly selecting its genes from the input geneset. In the case of the Random LGN the values of the performance parameters were very small, demonstrating that PC-IM realizes that the LGN is not a real LGN. This test shows the robustness of the method and that its performances depend on the type of LGN.

As last evaluation we have estimated the validity of the intrinsic performance evaluation and if the cut-off frequency intrinsically estimated is a good parameter to select the final gene expansion list. The results show that the cut-off frequency allows for selecting the gene list that is the best compromise between PPV and Se. In the absence of experimental validations is not possible to project, with certainty, this intrinsic estimate of PPV and Se to the expansion gene list.

The comparison between PC-IM and GENIES shows that GENIES gives the best expansion performances within the LGN, but it does find almost no extra genes. Moreover in GENIES, to get to the best performances is necessary to test different combinations between kernel matrix and algorithms (`http://www.genome.jp./tools/genies/`). From these considerations is evident that GENIES can not be used to expand a LGN with extra genes and that GENIES is not easy to be used by a user with little informatic knowledge.

PCL filename	**reference**	**number of conditions**	**tags**
GSE7820_set0_family	[Medintz et al., 2007]	6	filamentous growth, nitrogen utilization, signaling
GSE7820_set1_family	[Medintz et al., 2007]	6	filamentous growth, nitrogen utilization, signaling
GSE7820_setA_family	[Medintz et al., 2007]	12	filamentous growth, nitrogen utilization, signaling
GSE8898_setA_family	[Jansen et al., 2005]	6	evolution, fermentation, respiration
GSE9232_setA_family	[Daran-Lapujade et al., 2007]	9	carbon utilization, fermentation, oxygen level alteration, respiration
2010.Bernstein00 HDACrpd3sin3hda1	[Bernstein et al., 2000]	6	chromatin organization, histone modification
2010.Yoshimoto02 Ca.flt.knn.avg	[Yoshimoto et al., 2002]	24	cellular ion homeostasis, chemica, stimulus, signaling
GSE10066_setA_family	[Abbott et al., 2008]	12	chemical stimulus, oxygen level alteration, stress
GSE10073_setA_family	[Agarwal et al., 2008]	6	chemical stimulus, cofactor metabolism
GSE10521_setA_family	[Azzouz et al., 2009]	25	carbon utilization, diauxic shift, fermentation, respiration
GSE1311_setA_family	[Singh et al., 2005]	21	stress
GSE1312_setA_family	[Singh et al., 2005]	21	stress
GSE4272_set0_family	[Auld et al., 2006]	17	proteolysis, transcription
GSE5027_setA_family	[Barbara et al., 2007]	12	carbon utilization, fermentation
GSE6018_set0_family	[Benton et al., 2006]	12	DNA damage stimulus
GSE6018_set1_family	[Benton et al., 2006]	13	radiation
GSE8335_set00_family	[Berry and Gasch, 2008]	8	osmotic stress, stress
GSE8335_set01_family	[Berry and Gasch, 2008]	8	oxidative stress, stress
GSE8335_set02_family	[Berry and Gasch, 2008]	8	oxidative stress, stress
GSE8335_set03_family	[Berry and Gasch, 2008]	4	osmotic stress, stress
GSE8335_set04_family	[Berry and Gasch, 2008]	4	oxidative stress, stress
GSE8335_set05_family	[Berry and Gasch, 2008]	8	oxidative stress, stress
GSE8335_set06_family	[Berry and Gasch, 2008]	6	osmotic stress, stress
GSE8335_set07_familyy	[Berry and Gasch, 2008]	5	heat shock, stress
GSE8335_set08_family	[Berry and Gasch, 2008]	5	osmotic stress, stress
GSE8335_set09_family	[Berry and Gasch, 2008]	5	heat shock, stress
GSE8335_set10_family	[Berry and Gasch, 2008]	5	stress
GSE8335_set11_family	[Berry and Gasch, 2008]	5	heat shock, stress
GSE8335_set12_family	[Berry and Gasch, 2008]	5	osmotic stress, stress
GSE8335_set13_family	[Berry and Gasch, 2008]	5	heat shock, stress
GSE8624_set0_family	[Aragon et al., 2008]	26	stationary phase maintenance
GSE8624_set1_family	[Aragon et al., 2008]	20	stationary phase maintenance
GSE8900_setA_family	[Aguilera et al., 2006]	18	carbon utilization, chemical stimulus|nitrogen utilization, respiration
GSE9376_setA_family	[Smith and Kruglyak, 2008]	30	carbon utilization, evolution

Table 4.16: **Description of SGD expression data** (`http://www.yeastgenome.org/download-data/expression`) and used to expand the LGN3.

name of combination	**kernel matrix**	**algorithm type**
GENIES_1	linear kernel	kernel matrix regression
GENIES_2	Gaussian RBF kernel	kernel matrix regression
GENIES_3	Exponential kernel	kernel matrix regression
GENIES_4	Polynomial kernel	kernel matrix regression
GENIES_5	Linear kernel	Penalize kernel matrix regression
GENIES_6	Linear kernel	em-algorithm
GENIES_7	Linear kernel	kernel canonical correlation analysis
GENIES_8	Gaussian RBF kernel	penalized kernel matrix regression
GENIES_9	Gaussian RBF kernel	em-algorithm
GENIES_10	Gaussian RBF kernel	kernel canonical correlation analysis
GENIES_11	exponential kernel	penalized kernel matrix regression
GENIES_12	exponential kernel	em-algorithm
GENIES_13	exponential kernel	kernel canonical correlation analysis
GENIES_14	polynomial kernel	penalized kernel matrix regression
GENIES_15	polynomial kernel	em-algorithm
GENIES_16	polynomial kernel	kernel canonical correlation analysis

Table 4.17: **Different combination of kernel matrix and algorithms of GENIES.**

Chapter 5

Expansion of the Local Gene Networks with PC-IM: two case studies.

Once tested that PC-IM has good ability in expanding a LGN, the method was applied in the expansion of two real LGNs of *Arabidopsis thaliana*:

- the flowering network AtFOS-GRN which was also used to test the PC-IM;
- the flavonoid network (AtFlavonoids)

5.1 The *Arabidopsis thaliana* Floral Organ Specification- Gene Regulatory Network

The first case study dealt with the expansion of the FOS-GRN of *Arabidopsis thaliana* (AtFOS-GRN). This network is formed by 15 genes and it is involved in flower development. Its complete description is presented in Section 4.2.1. The parameters used for PC-IM in this run were:

- *tiles* size $t = 1000$;
- iteration number $i = 100$;
- subLGNs size $d = 10$;
- gene expression data reported in Section 4.2.2.

PC-IM produced a list of 314 genes expanding FOS-GRN. At this step the validation of the expansion was performed by an exhaustive bibliographic search. The genes, of the final output list, were assigned to four different classes, as described in Section 4.3.5:

- Class 1: genes related to flowering;

- Class 2: genes not directly related with flowering, but related to genes of Class 1;

- Class 3: genes unrelated with flowering;

- Class 4: genes not supported by references.

The complete expansion genes list of FOS-GRN is detailed in Table 5.2. Genes are ordered by ranking. The ranking is obtained considering how many times a gene is present in the output list of each iteration. The great number of genes of Class 4, indicates that the knowledge in the literature does not cover all the genes in the expanded FOS-GRN. It is worth to mention that on other hand, these genes of Class 4 represented a new potential source of knowledge.

5.1.1 PC-IM Output *versus* Random Output

Since the exhaustive bibliographic search did not completely cover the whole expansion list given by PC-IM, we included another test to evaluate the PC-IM results. This test tries to answer the question: "Is the performance of PC-IM in terms of PPV and e) similar to that observed using a list of randomly selected genes as Random Output? "

The Random Output has 314 genes, namely the same number of genes of the PC-IM Output. The 314 genes have been obtained by applying the method on randomly selected genes from ATH1 GeneChip. The analysis was repeated 10 times, thus giving 10 Random Outputs.

To compare the two outputs we used as reference (Gold Standard) a set of genes involved in flower development, but not included in FOS-GRN. This set of genes was obtained running the ANAP tool [Wang et al., 2012] on the FOS-GRN. ANAP tool is a platform of *Arabidopsis thaliana* by which is possible to select different databases or different interaction detection methods to generate networks [Wang et al., 2012]. To get our Gold Standard we have selected only the interaction detection methods. Starting from FOS-GRN it was possible to obtain different Gold Standard combining various methods by the drop-down menu called network filtering. The first Gold Standard (Gold Standard A) of 151 genes was obtained by selecting all entries present in the network filtering (100 interaction detection methods). In the second Gold Standard (Gold Standard B) there are only 36 genes and it was obtained by selecting only the ”hybridization experiment” entries in the network filtering. These hybridization experiments correspond to three Interaction Detection Methods. The hybridization screenings are based on molecular biology techniques used to discover interactions between molecules (e.g. protein-protein and DNA-protein interactions). The Gold Standard B was generated starting from *in vivo* experiments and therefore should be more realistic compared to Gold Standard A.

The PC-IM Output and Random Output were compared by a Likelihood Ratio (LR) test, typically used to compare the fit of two models, the null model (Gold Standard in our case) and the alternative model (the two different Output in our case). It is used in the diagnostic field to assess how good a diagnostic test (pre-test) is and to help in selecting an appropriate diagnostic tests or sequence of tests. The Positive LR (LR+) indicates the probability to have a positive result in a diseased subject respect to the same probability in a healthy subject:

$$LR+ = \frac{sensitivity}{(1 - specificity)} \tag{5.1}$$

If LR+ is positive and higher than 1.0 the probability that the diagnostic test is correct increases. If LR+ is equal to 1.0 the diagnostic test does not give any information on the disease prediction.

	Gold Standard A	**Gold Standard B**
PC-IM Outputs	8.58	10.23
Random Output	1.06	0.81

Table 5.1: **Comparison of the LR+ value of PC-IM (314 PC-IM genes) and the LR+ value of Random Output genes (314 random genes).**

In Table 5.1 the results of the LR+ test are showed. It is interesting to note that LR+ values of the PC-IM Output are much higher than 1.0, while LR+ values of the Random Output are very close to 1.0. This means that PC-IM results gave a gene expansion list that has a good probability to be correct. Respect to Gold standard A, with gene lists of the Gold Standard B, the LR+ of the PC-IM Output increases, while the LR+ of Random Output decreases. This analysis shows that the expansion of list FOS-GRN given by PC-IM is significantly not the same as if was generated by a randomly.

***Table 5.2**: 314 expansion genes of FOS-GRN.*

rnk	**AffyID**	**GeneID**	**Description**	**Class**	**Reference**
1	245819_at	AT1G26310	CAL (CAULIFLOWER)	1	[Grandi et al., 2012]
2	254391_at	AT4G21590	ENDO3 (ENDONUCLEASE 3)	1	[Gómez-Mena et al., 2005]; [Schmid et al., 2003]
3	259089_at	AT3G04960	similar to unknown protein [Arabidopsis thaliana] (TAIR: AT4G27980.1)	1	[Gómez-Mena et al., 2005]

Table 5.2: 314 expansion genes of FOS-GRN.

4	260355_at	AT1G69180	CRC (CRABS CLAW)	1	[Lee et al., 2005]
5	266505_at	AT2G47830	cation efflux family protein / metal tolerance protein, putative (MTPc1)	2	expressed in carpel (TAIR)
6	264041_at	AT2G03710	SEP4 (SEPALLATA4)	1	[Ditta et al., 2004]
7	267460_at	AT2G33810	SPL3 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 3)	1	[Deng et al., 2011]; [Yamaguchi et al., 2009]
8	250467_at	AT5G10100	TPPI (TREHALOSE-6- PHPSPHATE PHOSPHATASE)	1	[Iturriaga et al., 2009]; [Li et al., 2008]
9	256780_at	AT3G13640	RNaseL inhibitor protein 1	2	[Van Leene et al., 2010]
10	264444_at	AT1G27360	SPL11 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 11)	1	[Chen et al., 2010]; [Yamaguchi et al., 2009]
11	255609_s_at	AT4G01180	XH/XS domain-containing protein	2	[Ausin et al., 2009]
12	267144_at	AT2G38110	ATGPAT6/GPAT6 (GLYCEROL-3- PHOSPHATE ACYLTRANSFERASE 6)	1	[Li et al., 2012]
13	266888_s_at	AT2G44750	TPK2 (THIAMIN PYROPHOSPHOKINASE 2)	2	[Ajjawi et al., 2007]
14	247956_at	AT5G56970	CKX3 (CYTOKININ OXIDASE 3)	1	[Holst et al., 2011]
15	261499_at	AT1G28430	CYP705A24 (cytochrome P450, family 705, subfamily A, polypeptide 24)	1	[Mizutani and Ohta, 2010]
16	262231_at	AT1G68740	EXS (ERD1/XPR1/SYG1) family protein PHO	3	-
17	251543_at	AT3G58770	similar to hypothetical protein [Vitis vinifera] (GB:CAN63610.1)	2	[Causier et al., 2010]
18	253258_at	AT4G34400	DNA binding / transcription factor	2	[Zhang et al., 2005]

Table 5.2: 314 expansion genes of FOS-GRN.

19	253488_at	AT4G31610	REM1 (REPRODUCTIVE MERISTEM 1)	1	[Ståldal et al., 2012]; [Swaminathan et al., 2008]
20	263605_at	AT2G16480	SWIB complex BAF60b domain-containing protein	3	-
21	247869_at	AT5G57520	ZFP2 (ZINC FINGER PROTEIN 2)	1	[Cai and Lashbrook, 2008]
22	262642_at	AT1G62690	unknown protein	3	-
23	253712_at	AT4G29330	DER1 (DERLIN-1)	3	-
24	261375_at	AT1G53160	SPL4 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 4)	1	[Wang et al., 2009]
25	255448_at	AT4G02810	similar to unknown protein [Arabidopsis thaliana] (TAIR: AT1G03170.1)	1	[Wahl et al., 2010]
26	249939_at	AT5G22430	similar to unknown protein [Arabidopsis thaliana] (TAIR: AT2G27385.1)	1	[Sliwinski et al., 2006]; [Zik and Irish, 2003]
27	250570_at	AT5G08170	ATAIH/EMB1873 (AGMATINE IMINOHYDROLASE)	3	-
28	257034_at	AT3G19184	DNA binding	4	-
29	249614_at	AT5G37300	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G38995.1)	3	-
30	264489_at	AT1G27370	SPL10 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 10)	1	[Fornara and Coupland, 2009]
31	257051_at	AT3G15270	SPL5 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 5)	1	[Wang et al., 2009]
32	264752_at	AT1G23010	LPR1 (LOW PHOSPHATE ROOT1)	3	-
33	247718_at	AT5G59310	LTP4 (LIPID TRANSFER PROTEIN 4);	3	-
34	266319_s_at	AT3G10280	fatty acid elongase 3-ketoacyl-CoA synthase	2	[Bach and Faure, 2010]

Table 5.2: 314 expansion genes of FOS-GRN.

35	260438_at	AT1G68290	ENDO 2 (ENDONUCLEASE 2)	2	[Yu et al., 2005]
36	248227_at	AT5G53820	Late embryogenesis abundant protein (LEA) family protein	2	[Bies-Etheve et al., 2008]
37	263277_at	AT2G14110	haloacid dehalogenase-like hydrolase domain-containing protein	4	-
38	256259_at	AT3G12460	DEDDy 3'-5' exonuclease domain-containing protein	3	-
39	266814_at	AT2G44910	HB-4 (homeobox-leucine zipper protein 4)	1	[Zhang et al., 2005]
40	248167_at	AT5G54530	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G61667.1)	4	-
41	260097_at	AT1G73220	AtOCT1 (ORGANIC CATION/CARNITINE TRANSPORTER1)	3	-
42	252175_at	AT3G50700	AtIDD2 (INDETERMINATE(ID)-DOMAIN 2)	1	[Seo et al., 2011]
43	254663_at	AT4G18290	KAT2 (K+ ATPase 2)	3	-
44	256597_at	AT3G28500	60S acidic ribosomal protein P2 (RPP2C)	2	[Wang et al., 2009]
45	251986_at	AT3G53310	REM16 (REPRODUCTIVE MERISTEM 16)	1	[Wynn et al., 2011]
46	247447_at	AT5G62730	proton-dependent oligopeptide transport (POT) family protein	4	-
47	265261_at	AT2G42990	GDSL-motif lipase/hydrolase family protein	2	[Shi et al., 2011b]
48	257943_at	AT3G21840	ASK7 (SKP1-LIKE 7)	3	-
49	259616_at	AT1G47960	C/VIF1 (Cell wall/Vacuolar Inhibitor of Fructosidase 1)	4	-
50	248752_at	AT5G47600	heat shock protein-related	4	-
51	253309_at	AT4G33790	acyl CoA reductase, putative	3	-
52	250982_at	AT5G03150	JKD (JACKDAW)	3	-
53	255730_at	AT1G25460	oxidoreductase family protein	4	-
54	261514_at	AT1G71870	MATE efflux family protein	4	-

Table 5.2: 314 expansion genes of FOS-GRN.

55	255956_at	AT1G22015	DD46 (putative beta-1,3-galactosyltransferase 5)	2	[Bemer et al., 2008]
56	267528_at	AT2G45650	AGL6 (AGAMOUS LIKE-6)	1	[Rijpkema et al., 2009]; [Hsu et al., 2003]
57	248496_at	AT5G50790	nodulin MtN3 family protein	2	[Wellmer et al., 2006]
58	258506_at	AT3G06520	agenet domain-containing protein	4	-
59	253266_s_at	AT4G34080	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G45260.1)V	4	-
60	251563_at	AT3G57880	C2 domain-containing protein	4	-
61	260787_at	AT1G06230	GTE4 (GLOBAL TRANSCRIPTION FACTOR GROUP E4)	3	-
62	262835_at	AT1G14660	ATNHX8 (Na+/H+ exchanger 8)	3	-
63	266855_at	AT2G26920	ubiquitin-associated (UBA)/TS-N domain-containing protein	4	-
64	245458_at	AT4G16970	kinase	4	-
65	257504_at	AT1G52250	dynein light chain type 1 family protein	4	-
66	247758_at	AT5G59120	ATSBT4.13; subtilase	2	[Tung et al., 2005]
67	261925_at	AT1G22540	proton dependent oligopeptide transport (POT) family protein	4	-
68	250684_at	AT5G06650	GIS2 (GLABROUS INFLORESCENCE STEMS 2)	3	-
69	259802_at	AT1G72260	THI2.1 (THIONIN 2.1)	3	-
70	263567_at	AT2G15440	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G67210.1)	4	-
71	265672_at	AT2G31980	cysteine proteinase inhibitor 2	4	-
72	257579_at	AT3G11000	similar to kelch repeat-containing protein	2	[Wang et al., 2009]
73	259133_at	AT3G05400	sugar transporter ERD6-like 12	4	-
74	254175_at	AT4G24050	short-chain dehydrogenase/reductase (SDR) family protein	4	-

Table 5.2: 314 expansion genes of FOS-GRN.

75	245485_at	AT4G16230	GDSL-motif lipase/hydrolase family protein	4	-
76	259421_at	AT1G13910	leucine-rich repeat family protein	4	-
77	260461_at	AT1G10980	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G61670.1)	4	-
78	246310_at	AT3G51895	SULTR3;1 (SULFATE TRANSPORTER 1)	2	[Zuber et al., 2010]
79	255957_at	AT1G22160	senescence-associated protein-related	4	-
80	264907_at	AT2G17280	phosphoglycerate mutase family protein	4	-
81	252495_at	AT3G46770	transcriptional factor B3 family protein	1	[Romanel et al., 2011]; [Romanel et al., 2009]
82	247403_at	AT5G62740	band 7 family protein	4	-
83	263011_at	AT1G23250	caleosin-related	4	-
84	245087_at	AT2G39830	zinc ion binding	4	-
85	249219_at	AT5G42400	ATXR7 (TRITHORAX-RELATED7)	1	[Berr et al., 2009]
86	254558_at	AT4G19185	nodulin MtN21	4	-
87	263886_at	AT2G36960	TKI1 (TSL-KINASE INTERACTING PROTEIN 1)	3	-
88	246962_s_at	AT5G24800	ATBZIP9/BZO2H2 (BASIC LEUCINE ZIPPER O2 HOMOLOG 2)	3	-
89	263546_at	AT2G21550	bifunctional dihydrofolate reductase-thymidylate synthase, putative	3	-
90	259809_at	AT1G49800	unknown protein	4	-
91	253151_at	AT4G35670	glycoside hydrolase family 28 protein	4	-
92	251991_at	AT3G53340	nuclear transcription factor Y subunit B-10	2	[Siefers et al., 2009]
93	247469_at	AT5G62165	AGL42 (AGAMOUS LIKE 42)	1	[Dorca-Fornell et al., 2011]

Table 5.2: 314 expansion genes of FOS-GRN.

94	248251_at	AT5G53220	similar to unnamed protein product [Vitis vinifera] (GB:CAO69343.1)	4	-
95	260733_at	AT1G17640	RNA recognition motif (RRM)-containing protein	4	-
96	256664_at	AT3G12040	DNA-3-methyladenine glycosylase (MAG)	4	-
97	245488_at	AT4G16270	peroxidase 40 (PER40) (P40)	2	[Cosio and Dunand, 2010]
98	245275_at	AT4G15210	ATBETA-AMY (BETA-AMYLASE) 5	2	[Wilson et al., 2005]
99	255906_at	AT1G17790	DNA-binding bromodomain-containing protein	4	-
100	265689_at	AT2G24310	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G47170.1)	4	-
101	246380_at	AT1G57750	CYP96A15/MAH1 (MID-CHAIN ALKANE HYDROXYLASE 1)	3	-
102	255644_at	AT4G00870	basic helix-loop-helix (bHLH) family protein	2	[Hu et al., 2003]
103	245031_at	AT2G26360	binding	4	-
104	245842_at	AT1G58430	RXF26	1	[Shi et al., 2011b]
105	262905_at	AT1G59730	ATH7 (thioredoxin H-type 7)	4	-
106	262680_at	AT1G75880	EXL1 (extracellular lipase 1)	2	[Updegraff et al., 2009]
107	262675_at	AT1G75930	EXL6 (extracellular lipase 6)	2	[Updegraff et al., 2009]
108	260024_at	AT1G30080	glycosyl hydrolase family 17 protein	4	-
109	251863_at	AT3G54870	MRH2 (morphogenesis of root hair 2)	3	-
110	262443_at	AT1G47655	Dof-type zinc finger domain-containing protein	2	[Kaufmann et al., 2009]
111	264137_at	AT1G78960	ATLUP2 (LUPeol synthase 2)	4	-
112	264180_at	AT1G02190	CER1 protein	1	[Gómez-Mena et al., 2005]

Table 5.2: 314 expansion genes of FOS-GRN.

113	248918_at	AT5G45890	SAG12 (SENESCENCE-ASSOCIATED GENE 12)	3	-
114	254791_at	AT4G12910	SCPL20 (serine carboxypeptidase-like 20)	2	[Huang et al., 2009]
115	253518_at	AT4G31400	N-acetyltransferase	2	[Jiang et al., 2010]
116	256116_at	AT1G16858	CPuORF55 (Conserved peptide upstream open reading frame 55)	4	-
117	255199_at	AT4G07390	PQ-loop repeat family protein / transmembrane family protein	2	[Xiang et al., 2011]
118	251560_at	AT3G57920	SPL15 (squamosa promoter-binding protein)	1	[Deng et al., 2011]
119	259334_at	AT3G03790	ankyrin repeat family protein	3	-
120	266922_s_at	AT2G45950	ASK20 (ARABIDOPSIS SKP1-LIKE 20)	2	[Zhao et al., 2003]
121	261068_at	AT1G07450	tropinone reductase, putative	3	-
122	263741_at	AT2G20620	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G29550.1)	4	-
123	265180_at	AT1G23590	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G23600.1)	4	-
124	252692_at	AT3G43960	cysteine proteinase, putative	4	-
125	255768_at	AT1G16705	p300/CBP acetyltransferase-related protein-related	1	[Han et al., 2006]
126	256239_at	AT3G12470	DEDDy 3'-5' exonuclease domain-containing protein	4	-
127	250475_at	AT5G10180	AST68 (Sulfate transporter 2.1)	3	-
128	261575_at	AT1G01130	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G47170.1)	4	-
129	249865_at	AT5G22820	binding	4	-
130	250977_at	AT5G03070	binding	4	-
131	258652_at	AT3G09910	AtRABC2b/AtRab18C (Arabidopsis Rab GTPase homolog C2b)	4	-

Table 5.2: 314 expansion genes of FOS-GRN.

132	266605_at	AT2G46020	ATBRM/BRM/CHR2 (ARABIDOPSIS THALIANA BRAHMA)	1	[Smaczniak et al., 2012]
133	264992_at	AT1G67300	hexose transporter, putative	4	-
134	265441_at	AT2G20870	cell wall protein precursor	1	[Cai et al., 2007]; [Maizel et al., 2005]
135	257944_at	AT3G21850	ASK9 (ARABIDOPSIS SKP1-LIKE 9)	2	[Takahashi et al., 2004]
136	248572_at	AT5G49800	similar to unnamed protein product [Vitis vinifera] (GB:CAO46940.1)	4	-
137	262874_at	AT1G65020	similar to unnamed protein product [Vitis vinifera] (GB:CAO62149.1)	4	-
138	249144_at	AT5G43270	SPL2 (SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 2)	2	[Usami et al., 2009]
139	266171_at	AT2G38880	ATHAP3/ATNF-YB1/HAP3/HAP3A (NUCLEAR FACTOR Y SUBUNIT B1)	1	[Cai et al., 2007]
140	248559_at	AT5G50012	CPuORF36 (Conserved peptide upstream open reading frame 36)	4	-
141	256654_at	AT3G18880	ribosomal protein S17 family protein	4	-
142	245571_at	AT4G14695	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G22310.1)	2	[Hu and Vick, 2003]
143	254667_at	AT4G18280	glycine-rich cell wall protein-related	4	-
144	257158_at	AT3G24360	enoyl-CoA hydratase/isomerase family protein	4	-
145	252200_at	AT3G50280	transferase family protein	4	-
146	256418_at	AT3G06160	transcriptional factor B3 family protein	2	[Romanel et al., 2009]
147	258082_at	AT3G25905	CLE27 (CLAVATA3/ ESR-RELATED 27)	1	[Jun et al., 2010]; [Wijeratne et al., 2007]

Table 5.2: 314 expansion genes of FOS-GRN.

148	248368_at	AT5G51950	glucose-methanol-choline (GMC) oxidoreductase family protein	4	-
149	264066_at	AT2G27880	AGO 5 (ARGONAUTE protein)	1	[Tucker et al., 2012]
150	267431_at	AT2G34870	MEE26 (maternal effect embryo arrest 26)	2	[Kinoshita et al., 2010]
151	249942_at	AT5G22300	NIT4 (NITRILASE 4)	4	-
152	262604_at	AT1G15060	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G73750.1)	4	-
153	255054_s_at	AT4G09740	ATGH9B14 (GLYCOSYL HYDROLASE 9B14)	2	[Kang et al., 2008]
154	256219_at	AT1G56260	similar to hypothetical protein OsI_024078 [Oryza sativa (indica cultivar-group)] (GB:EAZ02846.1)	3	-
155	257118_at	AT3G20180	metal ion binding	4	-
156	255345_at	AT4G04460	aspartyl protease family protein	4	-
157	248022_at	AT5G56510	APUM12 (PUMILIO 12); RNA binding	4	-
158	254197_at	AT4G24040	ATTRE1/TRE1 (TREHALASE 1)	1	[Müller et al., 2001]
159	250491_at	AT5G09780	transcriptional factor B3 family protein	2	[Romanel et al., 2009]
160	250288_at	AT5G13350	auxin-responsive GH3 family protein	4	-
161	259382_s_at	AT3G16430	jacalin lectin family protein	3	-
162	252128_at	AT3G50870	MNP (MONOPOLE); transcription factor	2	[Zhang et al., 2005]
163	247747_at	AT5G59000	zinc finger (C3HC4-type RING finger) family protein	2	[Wang et al., 2009]
164	248985_at	AT5G45160	root hair defective 3 GTP-binding (RHD3) family protein	3	-
165	253108_at	AT4G35900	FD	1	[Wigge et al., 2005]
166	267481_at	AT2G02780	leucine-rich repeat transmembrane protein kinase, putative	3	-

Table 5.2: 314 expansion genes of FOS-GRN.

167	256381_at	AT1G66850	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	2	[Dou et al., 2011]
168	248941_s_at	AT5G45460	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G45470.1)	4	-
169	257402_at	AT1G23570	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G23580.1)	2	[Yang et al., 2007]
170	266772_s_at	AT4G16540	heat shock protein-related	4	-
171	247041_at	AT5G67180	AP2 domain-containing transcription factor		[Deng et al., 2011]
172	246045_at	AT5G19430	zinc finger (C3HC4-type RING finger) family protein	4	-
173	257509_at	AT1G63190	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G63200.1)	4	-
174	259294_at	AT3G05330	hypotetical protein	4	-
175	266073_at	AT2G18770	signal recognition particle binding	4	-
176	263738_at	AT1G60060	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G53900.2)	4	-
177	257124_at	AT3G20040	ATHXK4; ATP binding / hexokinase	4	-
178	266190_at	AT2G38840	guanylate-binding family protein	4	-
179	251635_at	AT3G57510	ADPG1 (endo-polygalacturonase 1)	2	[Shi et al., 2011b]; [Qi et al., 2011b]
180	261511_at	AT1G71770	PAB5 (POLY(A)-BINDING PROTEIN)	2	[Yang et al., 2007]
181	259054_at	AT3G03480	CHAT (ACETYL COA:(Z)-3-HEXEN-1-OL ACETYLTRANSFERASE)	4	-
182	266139_at	AT2G28085	auxin-responsive family protein	4	-
183	245349_at	AT4G16690	esterase/lipase/thioesterase family protein	2	[Yang et al., 2008]

Table 5.2: 314 expansion genes of FOS-GRN.

184	265107_s_at	AT1G63380	short-chain dehydrogenase/reductase (SDR) family protein	4	-
185	264481_at	AT1G77200	AP2 domain-containing transcription factor TINY, putative	4	-
186	246025_at	AT5G21150	Argonaute family protein 9	2	[Olmedo-Monfil et al., 2010]
187	249005_at	AT5G44630	terpene synthase/cyclase family protein	1	[Tholl et al., 2005]
188	254494_at	AT4G20050	QRT3 (QUARTET 3)	2	[Kang et al., 2008]
189	254234_at	AT4G23680	major latex protein-related / MLP-related	4	-
190	264342_at	AT1G12080	contains domain PTHR22683 (PTHR22683)	2	[Alves-Ferreira et al., 2007]
191	250636_at	AT5G07520	GRP18 (Glycine rich protein 18)	2	[Alves-Ferreira et al., 2007]
192	256783_at	AT3G13670	protein kinase family protein	4	-
193	261919_at	AT1G65980	TPX1 (THIOREDOXIN-DEPENDENT PEROXIDASE 1)	4	-
194	260374_at	AT1G73960	TAF2 (TBP-ASSOCIATED FACTOR 2)	2	[Mougiou et al., 2012]
195	246513_at	AT5G15680	binding	4	-
196	259799_at	AT1G72290	trypsin and protease inhibitor family protein / Kunitz family protein	2	[Bektas et al., 2012]
197	263092_at	AT2G16210	transcriptional factor B3 family protein	2	[Wijeratne et al., 2007]
198	245371_at	AT4G15750	invertase/pectin methylesterase inhibitor family protein	2	[Ma et al., 2012]
199	245769_at	AT1G30220	ATINT2 (INOSITOL TRANSPORTER 2)	2	[Aluri and Büttner, 2007]
200	258116_at	AT3G14520	terpene synthase/cyclase family protein	1	[Ro et al., 2006]
201	264204_at	AT1G22710	SUC2 (SUCROSE-PROTON SYMPORTER 2)	1	[Corbesier et al., 2007]

Table 5.2: 314 expansion genes of FOS-GRN.

202	249687_at	AT5G36150	ATPEN3 (PUTATIVE PENTACYCLIC TRITERPENE SYNTHASE 3)	1	[Posé et al., 2011]
203	259533_at	AT1G12530	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G56420.1)	4	-
204	256293_at	AT1G69440	AGO7 (ARGONAUTE7)	1	[Tantikanjana et al., 2009]
205	248883_at	AT5G46190	KH domain-containing protein	4	-
206	249020_at	AT5G44800	CHR4/MI-2-LIKE (chromatin remodeling 4)	1	[Smaczniak et al., 2012]
207	257011_at	AT3G14070	CAX9 (CATION EXCHANGER 9)	3	-
208	249379_at	AT5G40460	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G27630.1)	4	-
209	260701_at	AT1G32330	ATHSFA1D (Heat Shock transcription FSctor A1D)	3	-
210	265531_at	AT2G06200	AtGRF6 (GROWTH-REGULATING FACTOR 6)	2	[Kaufmann et al., 2009]
211	248073_at	AT5G55720	pectate lyase family protein	2	[Maizel et al., 2005]
212	260241_at	AT1G63710	CYP86A7 (cytochrome P450, family 86, subfamily A, polypeptide 7)	2	[Maizel et al., 2005]
213	254028_s_at	AT4G25850	oxysterol-binding family protein	4	-
214	257129_at	AT3G20100	CYP705A19 (cytochrome P450, family 705, subfamily A, polypeptide 19)	4	-
215	264500_at	AT1G09370	enzyme inhibitor/ pectinesterase	4	-
216	265151_at	AT1G51340	MATE efflux family protein	4	-
217	251696_at	AT3G56590	putative protein	4	-
218	254287_at	AT4G22960	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G11860.1)	4	-

Table 5.2: 314 expansion genes of FOS-GRN.

219	261726_at	AT1G76270	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G20550.1)	4	-
220	248642_at	AT5G49120	senescence-associated protein-related	4	-
221	259972_at	AT1G76420	CUC3 (CUP SHAPED COTYLEDON3)	1	[Li et al., 2010]
222	248246_at	AT5G53200	TRY (TRIPTYCHON)	4	-
223	262150_at	AT1G52520	FRS6 (FAR1-related sequence 6)	1	[Lin and Wang, 2004]
224	263869_at	AT2G22000	PROPEP6 (Elicitor peptide 6 precursor)	4	-
225	264830_at	AT1G03710	cysteine protease inhibitor	4	-
226	249103_at	AT5G43600	N-carbamyl-L-amino acid hydrolase, putative	4	-
227	254573_at	AT4G19420	pectinacetylesterase family protein	4	-
228	254574_at	AT4G19430	unknown protein	2	[Wang et al., 2009]
229	262122_at	AT1G02790	PGA4 (POLYGALACTURONASE 4)	4	-
230	264016_at	AT2G21220	auxin-responsive protein, putative	4	-
231	246250_at	AT4G36880	CP1 (CYSTEINE PROTEINASE1)	4	-
232	258488_at	AT3G02420	similar to hypothetical protein [Cleome spinosa] (GB:ABD96906.1)	4	-
233	260876_at	AT1G21460	nodulin MtN3 family protein	2	[Wellmer et al., 2006]
234	261150_at	AT1G19640	JMT (JASMONIC ACID CARBOXYL METHYLTRANSFERASE)	4	-
235	246312_at	AT1G31930	XLG3 (EXTRA-LARGE GTP-BINDING PROTEIN 3)	3	-
236	259659_at	AT1G55170	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G14750.1)	4	-

Table 5.2: 314 expansion genes of FOS-GRN.

237	263386_at	AT2G40150	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G55990.1)	4	-
238	255014_at	AT4G09960	STK (SEEDSTICK)	1	[Favaro et al., 2003]
239	259221_s_at	AT3G03530	NPC4 (NONSPECIFIC PHOSPHOLIPASE C4)	3	-
240	250630_at	AT5G07400	FHA (forkhead-associated)	1	[Koornneef et al., 1998]
241	256128_at	AT1G18140	LAC1 (Laccase 1)	4	-
242	265943_at	AT2G19570	CDA1 (CYTIDINE DEAMINASE 1)	4	-
243	247717_at	AT5G59320	LTP3 (LIPID TRANSFER PROTEIN 3)	3	-
244	257679_at	AT3G20470	encodes a glycine-rich protein that is expressed more abundantly in immature seed pods than in stems and leaves. Expression is not detected in roots or flowers	2	[Mangeon et al., 2010]
245	256286_at	AT3G12180	cornichon family protein	4	-
246	246601_at	AT1G31710	copper amine oxidase, putative	3	-
247	248111_at	AT5G55330	membrane bound O-acyl transferase (MBOAT) family protein / wax synthase-related	4	-
248	253987_at	AT4G26270	phosphofructokinase family protein	4	-
249	252142_at	AT3G51120	zinc finger (CCCH-type) family protein	4	-
250	256149_at	AT1G55110	ATIDD7 (ARABIDOPSIS THALIANA INDETERMINATE(ID)-DOMAIN 7)	4	-
251	260540_at	AT2G43500	RWP-RK domain-containing protein	3	-
252	260006_at	AT1G68000	ATPIS1 (Arabidopsis thaliana phosphatidylinositol synthase 1)	3	-

Table 5.2: 314 expansion genes of FOS-GRN.

253	250408_at	AT5G10930	CIPK5 (CBL-INTERACTING PROTEIN KINASE 5)	4	-
254	256743_at	AT3G29370	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G39240.1)	4	-
255	254737_at	AT4G13840	transferase family protein	4	-
256	262278_at	AT1G68640	PAN (PERIANTHIA)	1	[Irish, 2010]
257	251448_at	AT3G59845	NADP-dependent oxidoreductase, putative	4	-
258	259673_at	AT1G77800	PHD finger family protein	4	-
259	259472_at	AT1G18910	protein binding / zinc ion binding	4	-
260	247154_at	AT5G65710	HSL2 (HAESA-LIKE 2)	2	[Shi et al., 2011a]
261	257701_at	AT3G12710	methyladenine glycosylase family protein	4	-
262	260921_at	AT1G21540	AMP-binding protein, putative	4	-
263	260203_at	AT1G52890	ANAC019 (Arabidopsis NAC domain containing protein 19)	4	-
264	250588_at	AT5G07660	structural maintenance of chromosomes (SMC) family protein	4	-
265	252469_at	AT3G46920	protein kinase family protein	4	-
266	264621_at	AT2G17700	protein kinase family protein	4	-
267	249611_at	AT5G37370	ATSRL1; binding	4	-
268	251979_at	AT3G53140	O-diphenol-O-methyl transferase, putative	4	-
269	245982_at	AT5G13170	nodulin MtN3 family protein	2	[Wellmer et al., 2006]
270	245307_at	AT4G16770	oxidoreductase, 2OG-Fe(II) oxygenase family protein	4	-
271	263443_at	AT2G28630	beta-ketoacyl-CoA synthase family protein	4	-
272	258932_at	AT3G10150	ATPAP16/PAP16 (purple acid phosphatase 16)	1	[Zhu et al., 2005]
273	262083_at	AT1G56100	pectinesterase inhibitor domain-containing protein	4	-
274	264934_at	AT1G11880	unknown protein	3	-

Table 5.2: 314 expansion genes of FOS-GRN.

275	264057_at	AT2G28550	RAP2.7/TOE1 (TARGET OF EAT1 1)	1	[Yant et al., 2010]; [Krizek et al., 2000]
276	264902_at	AT1G23060	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G70950.1)	4	-
277	265166_at	AT1G23640	pseudogene, hypothetical protein, contains Pfam profile PF02713: Domain of unknown function DUF220	4	-
278	255822_at	AT2G40610	ATEXPA8 (ARABIDOPSIS THALIANA EXPANSIN A8)	4	-
279	247068_at	AT5G66800	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G50640.1)	4	-
280	262275_at	AT1G68710	haloacid dehalogenase-like hydrolase family protein	4	-
281	250732_at	AT5G06480	MD-2-related lipid recognition domain-containing protein / ML domain-containing protein	4	-
282	249758_at	AT5G24350	similar to unnamed protein product [Vitis vinifera] (GB:CAO48609.1)	4	-
283	253861_at	AT4G27680	MSP1 protein	1	[Nonomura et al., 2003]
284	256777_at	AT3G13780	similar to SMAD/FHA [Medicago truncatula] (GB:ABN05826.1)	4	-
285	253818_at	AT4G28330	ATP binding / ATPase, coupled to transmembrane movement of substances	4	-
286	263306_at	AT2G12480	SCPL43; serine carboxypeptidase	4	-
287	262549_at	AT1G31290	PAZ domain-containing protein / piwi domain-containing protein	2	[Liu et al., 2011]
288	247560_at	AT5G61090	proline-rich family protein	4	-
289	264019_at	AT2G21130	peptidyl-prolyl cis-trans isomerase / cyclophilin (CYP2) / rotamase	4	-

Table 5.2: 314 expansion genes of FOS-GRN.

290	256254_at	AT3G11290	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G19220.1)	4	-
291	248553_at	AT5G50170	C2 domain-containing protein / GRAM domain-containing protein	4	-
292	248442_at	AT5G51280	DEAD-box protein abstrakt, putative	3	-
293	257089_at	AT3G20520	glycerophosphoryl diester phosphodiesterase family protein	4	-
294	248763_at	AT5G47550	cysteine protease inhibitor, putative / cystatin, putative	4	-
295	255057_at	AT4G09840	unknown protein	4	-
296	262728_at	AT1G75820	CLV1 (CLAVATA 1)	1	[Lenhard and Laux, 2003]; [Clark et al., 1993]
297	248154_at	AT5G54400	methyltransferase	4	-
298	257299_at	AT3G28050	nodulin MtN21 family protein	4	-
299	264367_at	AT1G03350	BSD domain-containing protein	4	-
300	264906_at	AT2G17270	mitochondrial substrate carrier family protein	4	-
301	263434_at	AT2G28610	PRS (PRESSED FLOWER)	1	[Matsumoto and Okada, 2001]
302	245521_at	AT4G15880	ESD4 (EARLY IN SHORT DAYS 4)	1	[Quesada et al., 2005]
303	248694_at	AT5G48340	similar to unnamed protein product [Vitis vinifera] (GB:CAO70880.1)	4	-
304	245676_at	AT1G56670	GDSL-motif lipase/hydrolase family protein	4	-
305	266512_at	AT2G47690	NADH-ubiquinone oxidoreductase-related	4	-
306	249761_at	AT5G23970	transferase family protein	4	-
307	254201_at	AT4G24130	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G56580.1)	2	[Maizel et al., 2005]

Table 5.2: 314 expansion genes of FOS-GRN.

308	248236_at	AT5G53870	plastocyanin-like domain-containing protein	4	-
309	253676_at	AT4G29570	cytidine deaminase, putative / cytidine aminohydrolase, putative	4	-
310	257925_at	AT3G23170	similar to unknown protein	4	-
311	264396_at	AT1G12050	fumarylacetoacetase, putative	4	-
312	265007_s_at	AT1G61563	RALFL8 (RALF-LIKE 8)	4	-
313	248069_at	AT5G55650	unknown protein	4	-
314	258346_at	AT3G22690	pentatricopeptide (PPR) repeat-containing protein	4	-

Table 5.2: **314 expansion genes of FOS-GRN.** [rnk = ranking]

5.2 The *Arabidopsis thaliana* flavonoid pathway(AtFlavonoids)

In the second case study, we chose the flavonoid biosynthesis pathway as LGN to be expanded with PC-IM . The flavonoid pathway is well studied in plants, because these compounds are numerous (alone are 4.5% of the plant metabolism) [Routaboul et al., 2012] [Harborne and Williams, 2000] and are involved in many physiological mechanisms. For example they are involved in flower and fruit color [Winkel-Shirley, 2001], in abiotic defense responses (as UV protection, water and cold stresses [Ryan et al., 2002]; [Winkel-Shirley, 2001]) and in the interactions between plant and other biological organisms (other plants, microbes and animals) [Harborne and Williams, 2000]. Although it has been studied and characterized in numerous plant species, we chose for the *Arabidopsis thaliana* expansion test since it is plant the model species. Several mutants of genes involved in flavonoid synthesis are also available [Routaboul et al., 2012].

The flavonoid pathway is presented in Figure 5.1. The LGN was defined selecting a subgroup of 21 genes represented by blue squares in Figure 5.1 and listed in Table 5.3. We did not choose all the genes present in Figure 5.1, in order to see if these genes were included with other new genes in the expansion gene list of PC-IM.

The PC-IM parameters adopted in the expansion of the LGN were:

- *tiles* size $t = 1000$;
- iteration number $i = 100$;
- subLGNs size $d = 14$;

Probe	GeneID	Abbreviation	Description
245126_at	AT2G47460	MYB12	Myb domani protein 12; DNA binding / transcription activator/ transcription factor
247354_at	AT5G63590	FLS	Flavonol synthase; flavonol synthase
248185_at	AT5G54060	UFGT	UDP-GLUCOSE:FLAVONOID 3-O-GLUCOSYL TRANSFERASE; transferase, transferring glycosyl groups
248200_at	AT5G54160	OMT	O-METHYLTRANSFERASE 1
249215_at	AT5G42800	DFR	DIHYDROFLAVONOL 4-REDUCTASE
249704_at	AT5G35550	TT2	Transparent Testa 2; DNA binding / transcription factor
249739_at	AT5G24520	TTG1	Transparent Testa Glabra1; nucleotide binding
249851_at	AT5G23260	TT16	Transparent Testa 16; transcription factor
250207_at	AT5G13930	CHS	CHALCONE SYNTHASE; TT4 (Transparent Testa 4); naringenin-chalcone synthase
250533_at	AT5G08640	FLS	FLAVONOL SYNTHASE
250558_at	AT5G07990	F3'H	flavonoid 3'-monooxygenase/ oxygen binding; TT7 (Transparent Testa 7)
251223_at	AT3G62610	MYB11	myb domain protein 11; DNA binding / transcription factor
251504_at	AT3G59030	MATE-TT12	Transparent Testa 12; antiporter/ solute:hydrogen antiporter/ transporter
251827_at	AT3G55120	CHI	chalcone-flavonone isomerase 1; A11/CFI/TT5 (Transparent Testa 5); chalcone isomerase
252123_at	AT3G51240	F3H	F3H (Transparent Testa 6); naringenin 3-dioxygenase
252534_at	AT3G46130	MYB111	myb domain protein 111; DNA binding / transcription factor
254283_s_at	AT4G22870; AT4G22880	LDOX	(leucoanthocyanidin dioxygenase), putative / anthocyanidin synthase, putative
255056_at	AT4G09820	TT8	TT8 (Transparent Testa 8); DNA binding / transcription factor
259751_at	AT1G71030	MYBL2	Arabidopsis myb-like 2; DNA binding / transcription factor
262528_at	AT1G17260	AHA10	AUTOINHIBITED H(+)-ATPASE ISOFORM 10; ATPase
264401_at	AT1G61720	ANR	BANYULS

Table 5.3: **Description of the genes of *Arabidopsis thaliana* flavonoids LGN.**

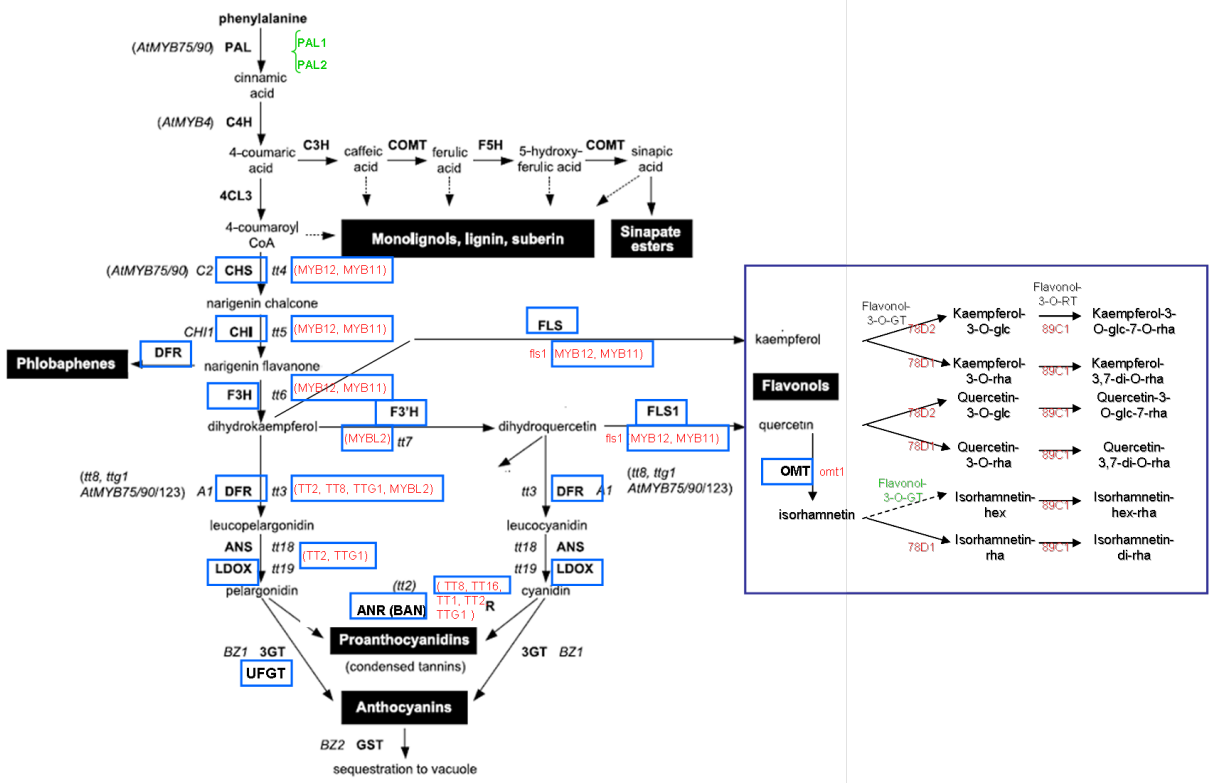


Figure 5.1: **Scheme of the phenylpropanoid biosynthetic pathway of *Arabidopsis thaliana*.**
The flavonoids pathway is part of phenylpropanoid pathway and it starts from the conversion of 4-coumaroyl-CoA in naringenin chalcone, leading o the production of three main classes of compounds, Flavonols, Proanthocyanidins and Anthocyanins (black boxes). Enzymes are indicated in bold upper-case letters and regulatory genes are indicated in parentheses and/or in red colour (Figure modified from Routaboul et al. [2012]).

- gene expression data are reported in Section 4.2.2.

The *tile* size was the same adopted in Section 4.3.1.

The value d was obtained maintaining the same ratio between the size of the LGN and the size of subLGN considered in the previous expansion (FOS-GRN).

PC-IM performances are summarized in Figure 5.2. The minimum value d_{min} is obtained with frequency values within the range (60-71)%. If d_{min} gives not an unique value of frequency, the selected cut-off value will be the mean of the range frequency values with the same d_{min} value. In the case of the AtFlavonoids LGN the cut-off frequency was: 65.5 %. The cut-off value is represented with black line in Figure 5.3-PPV and Se curve. Values of PPV and Se were 59.00 % and 26.00 %, respectively at the selected cut-off frequency.

PC-IM gave a final expansion genes list of 382 genes. To evaluate if the 382 genes were related to the flavonoid pathway, a bibliographic search was done (Table 5.4). The classification of the 382 genes followed the same criteria presented in Section 4.3.5 (Figure 5.3).

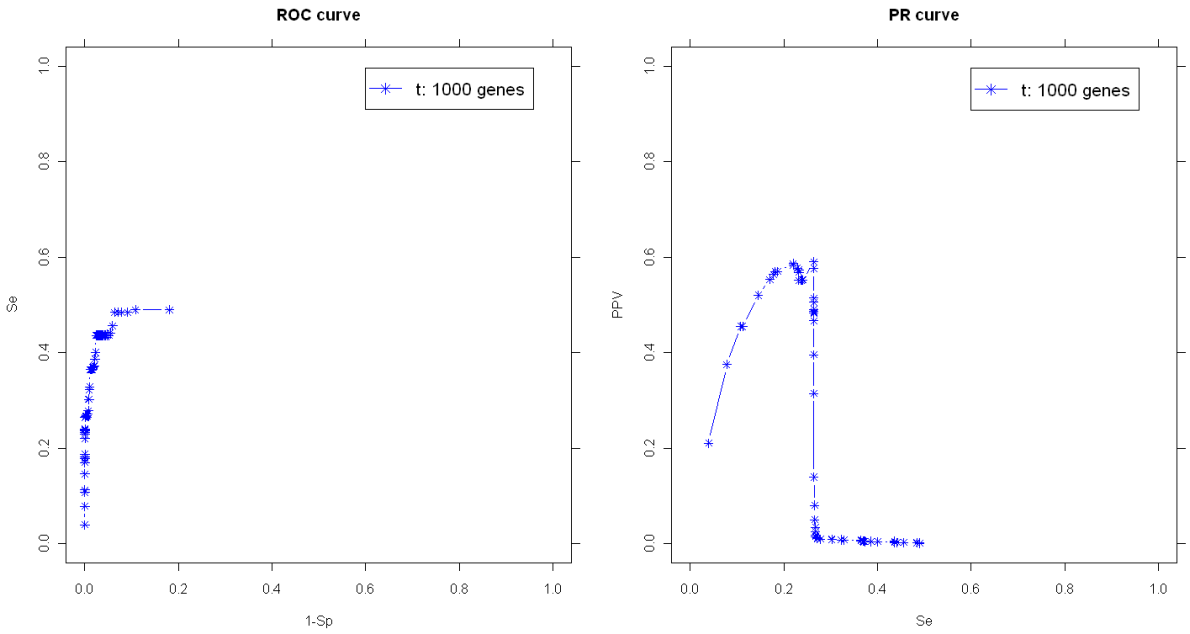


Figure 5.2: **ROC curve and PR curve of the phenylpropanoid pathway.**

5.3 Discussion

In this Chapter two different LGNs of *Arabidopsis thaliana* were expanded with PC-IM. The first LGN (FOS-GRN) regards flower development, while the second LGN (At-Flavonoids) is a subnetwork of the flavonoid pathway.

The parameters used in the PC-IM runs and the gene expression data, give as input, were the same for both LGNs.

The intrinsic performances of the FOS-GRN expansion (PPV= 82.23 %, Se = 46.70 %, Figure 4.7-PPV and Se curve) are greater than those of AtFlavonoids LGN (PPV = 59.00 %, Se = 26.00 %, Figure 5.3-PPV and Se curve). The lower value of PPV and Se of the AtFlavonoids expansion can be explained on the basis of the gene expression data used: among the 393 hybridization experiments, 40 concerned flower development (AT4, Section 4.2.2), while no gene expression data specific for the flavonoid pathway were present. It is worth to mention that twenty hybridization experiments were related

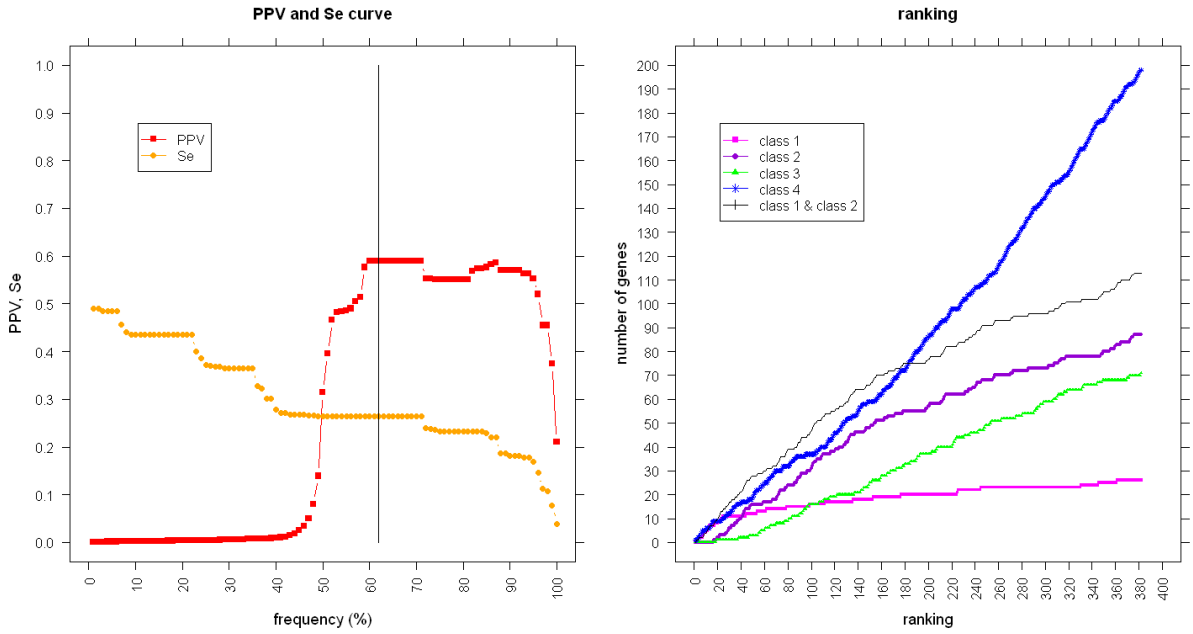


Figure 5.3: **PPV-Se curve and ranking curve of the flavonoids expansion.**

to plant defense responses (AT13, 4.2.2) and it is known that flavonoids are also involved in plant defense response [Ryan et al., 2002]; [Winkel-Shirley, 2001].

In both expansion gene lists we can distinguish the top ranking genes from the other.

- **top ranking genes** (higher frequency value). For these genes the values of PPV and Se are higher respect to those estimated by PC-IM during the intrinsic performance assessment. Indeed the number of genes belonging to Class 1 and Class 2 is much greater than that of genes Class 3 and Class 4 (Figure 4.7-ranking, Figure 5.3-ranking);

- **other genes** (lower frequency values). In this case the number of genes obtained combining Class1 and Class 2 is bigger than the number of genes of Class 3, but not of Class 4. This means the expansion genes provided by PC-IM are indeed related to the LGN, but knowledge is missing to finally validate them.

PC-IM gives as output a list of hypothetical candidate genes to the expansion of a specific LGN. This is the starting point to design an *in vivo* experiment to evaluate the real connection of the new genes with the LGN. In general, the *in vivo* experiments are expensive, both in terms of time and costs. Starting from this consideration, PC-IM gives an intrinsic evaluation performance to estimate the goodness of the PC-IM results. To support the reliability of the expanded gene list, the estimated performance

is done with a cautionary approach. The cautionary term indicates that the estimate of the intrinsic performances is made considering as FP all the genes found by PC-IM and not included in the LGN and as TP only genes belonging to LGN (Chapter 3-Step4: intrinsic performances assessment). This criterium, has a major effect on PPV calculation (Formule 3.1 and Formule 3.4).

Aim of this Chapter was to validate the output of PC-IM (expansion gene list) and to understand if the intrinsic evaluation performance is a good parameter to be used for selecting the expansion gene list. Unfortunately, the bibliographic search can not be the sole criterium for having a definitive validation due to the lack of information and to the specificity of some articles. The insufficient availability of information is underlined by the high number of genes in Class 4. About the specificity issues, it might well be that for genes belonging to Class 3, there are not yet specific studies linking them to the LGN. This implies that *in vivo* experiments should be planned to test the real involvement of a gene in the expanded LGN.

***Table 5.1**: Description of 382 expansion genes of the flavonoids pathway.*

rnk	AffyID	locus	annotation	Class	reference
1	259844_at	AT1G73560	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	4	-
2	253276_at	AT4G34050	caffeoyl-CoA 3-O-methyltransferase, putative	1	[Do et al., 2007]; [Besseau et al., 2007]
3	262083_at	AT1G56100	pectinesterase inhibitor domain-containing protein	4	-
4	263845_at	AT2G37040	PAL1 (PHE AMMONIA LYASE 1)	1	[Olsen et al., 2008]
5	248365_at	AT5G52500	similar to unknown protein	4	-
6	262545_at	At1g31250	proline-rich family protein	4	-
7	264500_at	AT1G09370	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	4	-
8	245624_at	AT4G14090	UDP-glucoronosyl/UDP-glucosyl transferase family protein	1	[Tohge et al., 2005]; [Gonzalez et al., 2007]
9	265091_s_at	AT1G03940	transferase family protein	1	[Tohge et al., 2005]
10	253724_at	AT4G29285	LCR24 (Low-molecular-weight cysteine-rich 24)	4	-
11	267620_at	AT2G39640	glycosyl hydrolase family 17 protein	1	[Marinova et al., 2007]
12	245560_at	AT4G15480	UGT84A1 (UDP-glycosyltransferase)	1	[Yonekura-Sakakibara et al., 2008]; [Stracke et al., 2007]
13	256924_at	AT3G29590	AT5MAT (anthocyanin 5-O-glucoside-O-malonyltransferase)	1	[D Auria et al., 2007]
14	257878_at	AT3G17150	plant invertase/pectin methylesterase inhibitor domain-containing protein	4	-
15	266572_at	AT2G23840	HNH endonuclease domain-containing protein	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

16	245892_at	AT5G09370	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	4	-
17	265248_at	AT2G43010	PIF4 (PHYTOCHROME INTERACTING FACTOR 4)	2	[Huq and Quail, 2002]
18	252958_at	AT4G38620	MYB4 (myb domain protein 4)	1	[Preston et al., 2004]; [Zhao et al., 2007]
19	263892_at	AT2G36890	ATMYB38/MYB38/RAX2 (myb domain protein 38)	3	-
20	258352_at	AT3G17600	IAA31 (auxin-responsive protein IAA31)	2	[Peer and Murphy, 2007]
21	249063_at	AT5G44110	POP1: ABC transporter I family member 21	1	[Molas et al., 2006]
22	267470_at	AT2G30490	ATC4H/C4H/CYP73A5 (CINNAMATE 4-HYDROXYLASE, CINNAMATE-4-HYDROXYLASE)	2	[Zhao et al., 2007]; [Besseau et al., 2007]
23	261181_at	AT1G34580	sugar transporter protein 5	1	[Tohge et al., 2005]
24	250251_at	AT5G13670	nodulin MtN21 family protein	4	-
25	250794_at	AT5G05270	chalcone-flavanone isomerase family protein	1	[Winkel-Shirley, 2002]
26	262396_at	AT1G49470	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G55230.1)	4	-
27	267262_at	AT2G22990	SNG1 (SINAPOYLGLUCOSE 1)	2	[Ruegger and Chapple, 2001]; [Fraser et al., 2007]
28	254786_at	AT4G12890	gamma interferon responsive lysosomal thiol reductase family protein / GILT family protein	4	-
29	256937_at	AT3G22620	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	2	[Hoang et al., 2012]
30	247262_at	AT5G64440	ATFAAH (ARABIDOPSIS THALIANA FATTY ACID AMIDE HYDROLASE)	2	[Thors et al., 2009]
31	247785_at	AT5G58820	subtilase family protein	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

32	249576_at	AT5G37690	GDSL-motif lipase/hydrolase family protein	2	[Riemann et al., 2008]
33	254474_at	AT4G20390	integral membrane family protein	4	-
34	260048_at	AT1G73750	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G15060.1)	4	-
35	263988_at	AT2G42830	SHP2 (SHATTERPROOF 2)	2	[Dardick et al., 2010]; [Scheible et al., 2004]
36	249061_at	AT5G44550	integral membrane family protein	4	-
37	261933_at	AT1G22410	2-dehydro-3-deoxyphosphoheptonate aldolase, putative	2	[Rohde et al., 2004]
38	254336_at	AT4G22050	aspartyl protease family protein	3	-
39	253580_at	AT4G30400	zinc finger (C3HC4-type RING finger) family protein	2	[Serrano and Guzmán, 2004]
40	253679_at	AT4G29610	cytidine deaminase, putative / cytidine aminohydrolase, putative	4	-
41	255403_at	AT4G03400	DFL2 (DWARF IN LIGHT 2)	2	[Takase et al., 2003]
42	245371_at	AT4G15750	invertase/pectin methylesterase inhibitor family protein	2	[Bolouri-Moghaddam et al., 2010]
43	264557_at	AT1G09550	pectinacetylesterase, putative	2	[Marín-Rodríguez et al., 2002]
44	266736_at	AT2G46960	CYP709B1 (cytochrome P450, family 709, subfamily B, polypeptide 1)	2	[Huang et al., 2006]
45	258590_at	AT3G04280	ARR22 (ARABIDOPSIS RESPONSE REGULATOR 22)	1	[Horák et al., 2008]; [Deikman and Hammer, 1995]
46	254394_at	AT4G21630	subtilase family protein	4	-
47	264214_s_at	AT1G65330	PHE1 (PHERES1)	3	-
48	257934_at	AT3G25420	SCPL21 (serine carboxypeptidase-like 21)	2	[Fraser et al., 2007]

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

49	266004_at	AT2G37330	ALS3 (ALUMINUM SENSITIVE 3)	2	[Larsen et al., 2004]
50	262589_s_at	AT1G15150	MATE efflux family protein	4	-
51	259107_at	AT3G05460	sporozoite surface protein-related	4	-
52	253699_at	AT4G29800	PLA IVD/PLP8 (Patatin-like protein 8)	4	-
53	248110_at	AT5G55320	membrane bound O-acyl transferase (MBOAT) family protein / wax synthase-related	4	-
54	250083_at	AT5G17220	ATGSTF12 (GLUTATHIONE S-TRANSFERASE 26)	1	[Tohge et al., 2005]
55	247697_at	AT5G59810	ATSBT5.4; subtilase	3	-
56	245501_at	AT4G15620	integral membrane family protein	4	-
57	255142_at	AT4G08390	SAPX; L-ascorbate peroxidase	3	-
58	246125_at	AT5G19875	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G31940.1)	4	-
59	250230_at	AT5G13900	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	2	[Hoang et al., 2012]
60	248593_at	AT5G49180	pectinesterase family protein	4	-
61	264078_at	AT2G28470	BGAL8 (BETA-GALACTOSIDASE 8)	3	-
62	245628_at	AT1G56650	PAP1 (PRODUCTION OF ANTHOCYANIN PIGMENT 1)	1	[Tohge et al., 2005]; [Broun, 2005]
63	264934_at	AT1G61090	hypothetical protein	4	-
64	246749_at	AT5G27830	similar to hypothetical protein [Vitis vinifera] (GB:CAN74239.1)	4	-
65	253204_at	AT4G34460	AGB1 (GTP BINDING PROTEIN BETA 1	3	-
66	254561_at	AT4G19160	binding	4	-
67	258457_at	AT3G22425	IGPD; imidazoleglycerol-phosphate dehydratase	2	[Glynn et al., 2005]

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

68	248260_at	AT5G53240	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G55270.1)	4	-
69	262549_at	AT1G31290	PAZ domain-containing protein / piwi domain-containing protein	4	-
70	248405_at	AT5G51480	SKS2 (SKU5 SIMILAR 2)	3	-
71	247515_at	AT5G61740	ATATH14 (ABC2 homolog 14)	2	[Morris and Zhang, 2006]
72	247747_at	AT5G59000	zinc finger (C3HC4-type RING finger) family protein	2	[Kosarev et al., 2002]
73	245204_at	AT5G12270	oxidoreductase, 2OG-Fe(II) oxygenase family protein	2	[Van Damme et al., 2008]
74	264898_at	AT1G23205	invertase/pectin methylesterase inhibitor family protein	2	[Zhang et al., 2007]
75	267218_at	AT2G02515	unknown protein	4	-
76	253186_at	AT4G35270	RWP-RK domain-containing protein	4	-
77	245734_at	AT1G73480	hydrolase, alpha/beta fold family protein	3	-
78	260599_at	AT1G55940	CYP708A1 (cytochrome P450, family 708, subfamily A, polypeptide 1)	2	[Huang et al., 2006]
79	265290_at	AT2G22590	glycosyltransferase family protein	1	[Stracke et al., 2007]
80	245090_at	AT2G40900	nodulin MtN21 family protein	2	[Ranocha et al., 2010]
81	253657_at	AT4G30110	HMA2 (Heavy metal ATPase 2)	3	-
82	257147_at	AT3G27270	similar to DNA-binding storekeeper protein-related [Arabidopsis thaliana] (TAIR:AT5G14280.1)	4	-
83	265355_at	AT2G16760	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G47370.1)	4	-
84	260873_at	AT1G21580	hydroxyproline-rich glycoprotein family protein	3	-
85	247463_at	AT5G62210	embryo-specific protein-related	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

86	245264_at	AT4G17245	zinc finger (C3HC4-type RING finger) family protein	2	[Gechev et al., 2008]
87	264610_at	AT1G04645	self-incompatibility protein-related	4	-
88	254146_at	AT4G24260	ATGH9A3/KOR3 (ARABIDOPSIS THALIANA GLYCOSYL HYDROLASE 9A3)	2	[Mølhøj et al., 2001]
89	259042_at	AT3G03450	RGL2 (RGA-LIKE 2)	2	[Lee et al., 2010]
90	261220_at	AT1G19970	ER lumen protein retaining receptor family protein	3	-
91	246161_at	AT5G20900	JAZ12/TIFY3B (JASMONATE-ZIM-DOMAIN PROTEIN 12)	2	[Qi et al., 2011a]
92	253135_at	AT4G35830	ACO1 (aconitate hydratase 1)	2	[Gupta et al., 2012]
93	265590_at	AT2G20160	MEO (MEIDOS); ubiquitin-protein ligase	3	-
94	254151_at	AT4G24390	F-box family protein (FBX14)	3	-
95	260761_at	AT1G49150	unknown protein	4	-
96	249497_at	AT5G39220	hydrolase, alpha/beta fold family protein	3	-
97	263156_at	AT1G54030	GDSL-motif lipase, putative	2	[Riemann et al., 2008]
98	244974_at	ATCG00700	PSII low MW protein	3	-
99	266625_at	AT2G35380	peroxidase 20 (PER20) (P20)	1	[Yamasaki et al., 1997]
100	245101_at	AT2G40890	CYP98A3 (cytochrome P450, family 98, subfamily A,polypeptide 3)	2	[Besseau et al., 2007]
101	257628_at	AT3G26290	CYP71B26 (cytochrome P450, family 71, subfamily B, polypeptide 26)	2	[Huang et al., 2006]
102	253277_at	AT4G34230	CAD5 (CINNAMYL ALCOHOL DEHYDROGENASE 5)	2	[Thévenin et al., 2011]; [Besseau et al., 2007]
103	265359_at	AT2G16720	MYB7 (myb domain protein 7)	2	[Causier et al., 2012]

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

104	256283_at	AT3G12540	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G39690.1)	4	-
105	245467_at	AT4G16610	zinc finger (C2H2 type) family protein	2	[Dinkins et al., 2012]
106	253127_at	AT4G36060	transcription factor bHLH11	4	-
107	251443_at	AT3G59940	kelch repeat-containing F-box family protein	3	-
108	248297_at	AT5G53100	oxidoreductase, putative	4	-
109	267198_at	AT2G30810	gibberellin-regulated family protein	2	[Herridge, 2012]
110	248619_at	AT5G49630	AAP6 (AMINO ACID PERMEASE 6)	2	[Hunt et al., 2010]
111	248791_at	AT5G47350	palmitoyl protein thioesterase family protein	3	-
112	261792_at	AT1G15950	CCR1 (CINNAMOYL COA REDUCTASE 1)	1	[Thévenin et al., 2011]; [Besseau et al., 2007]
113	260389_at	AT1G74055	unknown protein	4	-
114	244999_at	ATCG00190	Chloroplast DNA-dependent RNA polymerase B subunit	4	-
115	245158_at	AT2G33130	RALFL18 (RALF-LIKE 18)	4	-
116	258760_at	AT3G10780	emp24/gp25L/p24 family protein	3	-
117	246966_at	AT5G24850	CRY3 (CRYPTOCHROME 3)	2	[Huang et al., 2006]; [Onda et al., 2008]
118	250022_at	AT5G18210	short-chain dehydrogenase/reductase (SDR) family protein	4	-
119	249502_s_at	AT5G39280	ATEXPA23 (EXPANSIN A23)	4	-
120	253064_at	AT4G37730	ATBZIP7 (BASIC LEUCINE-ZIPPER 7)	4	-
121	255357_at	AT4G03930	pectinesterase 42	2	[Chen et al., 2011]
122	253956_at	AT4G26700	ATFIM1 (fimbrin 1)	3	-
123	263473_at	AT2G31750	UGT74D1 (UDP-GLUCOSYL TRANSFERASE 74D1)	4	-
124	255127_at	AT4G08300	nodulin MtN21 family protein	2	[Ranocha et al., 2010]

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

125	266196_at	AT2G39110	protein kinase, putative	4	-
126	256773_at	AT3G13630	unknown protein	4	-
127	261271_at	AT1G26795	self-incompatibility protein-related	4	-
128	261609_at	AT1G49740	phospholipase C	2	[Agullo et al., 1997]
129	246087_at	AT5G20580	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G06005.1)	4	-
130	251968_at	AT3G53100	GDSL-motif lipase/hydrolase family protein	2	[Riemann et al., 2008]
131	263982_at	AT2G42860	unknown protein	4	-
132	262989_at	AT1G23420	INO (INNER NO OUTER)	2	[Gallagher and Gasser, 2008]
133	256528_at	AT1G66140	ZFP4 (ZINC FINGER PROTEIN 4)	2	[Causier et al., 2012]
134	262516_at	AT1G17190	ATGSTU26 (Glutathione S-transferase (class tau) 26)	2	[Nutricati et al., 2006]
135	266386_at	AT2G32370	DNA binding / transcription factor	4	-
136	253017_at	AT4G37970	mannitol dehydrogenase, putative	1	[Kim et al., 2007]
137	255691_at	AT4G00370	ANTR2 (anion transporter 2)	3	-
138	265253_at	AT2G02020	proton-dependent oligopeptide transport (POT) family protein	2	[Weichert et al., 2012]
139	251071_at	AT5G01950	ATP binding / kinase/ protein serine/threonine kinase	4	-
140	245305_at	AT4G17215	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G47635.1)	4	-
141	249567_at	AT5G38020	S-adenosyl-L-methionine:carboxyl methyltransferase family protein	4	-
142	244904_at	ATMG00670	hypothetical protein	4	-
143	255777_at	AT1G18630	GR-RBP6 (glycine-rich RNA-binding protein 6)	3	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

144	250907_at	AT5G03670	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G36420.1)	4	-
145	253890_s_at	AT4G27585	band 7 family protein	3	-
146	262432_at	AT1G47530	ripening-responsive protein, putative	2	[Thompson et al., 2010]
147	254791_at	AT4G12910	SCPL20 (serine carboxypeptidase-like 20)	2	[Fraser et al., 2007]; [Floerl et al., 2012]
148	254965_at	AT4G11090	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G23790.1)	4	-
149	260948_at	AT1G06100	fatty acid desaturase family protein	3	-
150	264247_at	AT1G60160	potassium transporter family protein	3	-
151	266118_at	AT2G02130	LCR68/PDF2.3 (Low-molecular-weight cysteine-rich 68)	2	[Siddique et al., 2011]
152	262726_at	AT1G43640	AtTLP5 (TUBBY LIKE PROTEIN 5)	3	-
153	265954_at	AT2G37260	TTG2 (TRANSPARENT TESTA GLABRA 2)	1	[Ishida et al., 2007]
154	256994_s_at	AT3G25830	ATTPS-CIN (TERPENE SYNTHASE-LIKE SEQUENCE-1,8-CINEOLE)	2	[Chen et al., 2004]
155	259576_at	AT1G35330	zinc finger (C3HC4-type RING finger) family protein	4	-
156	265846_at	AT2G35770	SCPL28 (serine carboxypeptidase-like 28)	2	[Fraser et al., 2007]
157	260066_at	AT1G73610	GDSL-motif lipase/hydrolase family protein	4	-
158	265224_at	AT2G36710	pectinesterase family protein	4	-
159	255281_at	AT4G04970	ATGSL1 (GLUCAN SYNTHASE LIKE-1)	3	-
160	260166_at	AT1G79840	GL2 (GLABRA 2); DNA binding / transcription factor	3	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

161	247696_at	AT5G59780	MYB59 (myb domain protein 59)	4	-
162	260851_at	AT1G21890	nodulin MtN21 family protein	4	-
163	261308_at	AT1G48480	RKL1 (Receptor-like kinase 1)	2	[Tarutani et al., 2004]
164	262744_at	AT1G28680	transferase family protein	4	-
165	248022_at	AT5G56510	APUM12 (ARABIDOPSIS PUMILIO 12)	3	-
166	249856_at	AT5G22980	SCPL47 (serine carboxypeptidase-like 47)	2	[Fraser et al., 2007]
167	250416_at	AT5G11220	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G64870.1)	4	-
168	257943_at	AT3G21840	ASK7 (ARABIDOPSIS SKP1-LIKE 7)	3	-
169	262503_at	AT1G21670	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G21680.1)	4	-
170	267361_at	AT2G39920	acid phosphatase class B family protein	4	-
171	255073_at	AT4G09090	glycosyl hydrolase family protein 17	4	-
172	248011_at	AT5G56300	GAMT2; S-adenosylmethionine-dependent methyltransferase	2	[Varbanova et al., 2007]
173	257944_at	AT3G21850	ASK9 (ARABIDOPSIS SKP1-LIKE 9)	3	-
174	247430_at	AT5G62610	basic helix-loop-helix (bHLH) family protein	4	-
175	253096_at	AT4G37330	CYP81D4 (cytochrome P450, family 81, subfamily D, polypeptide 4)	4	-
176	257855_at	AT3G13040	myb family transcription factor	4	-
177	256186_at	AT1G51680	4CL1 (4-COUMARATE:COA LIGASE 1)	1	[Harding et al., 2002]
178	259871_at	AT1G76800	nodulin, putative	3	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

179	246627_s_at	AT2G45300	3-phosphoshikimate 1-carboxyvinyltransferase / 5-enolpyruvylshikimate-3-phosphate / EPSP synthase	2	[Chen et al., 2006]; [Logemann et al., 2000]
180	260913_at	AT1G02500	SAM1 (S-adenosylmethionine synthetase 1)	3	-
181	247568_at	AT5G61260	chromosome scaffold protein-related	4	-
182	263638_at	AT2G25310	carbohydrate binding	4	-
183	260990_at	AT1G12180	similar to heat shock protein-related [Arabidopsis thaliana] (TAIR:AT5G47600.1)	4	-
184	260753_at	AT1G49230	zinc finger (C3HC4-type RING finger) family protein	4	-
185	254564_at	AT4G19170	NCED4 (NINE-CIS-EPOXYCAROTENOID DIOXYGENASE 4)	3	-
186	258503_at	AT3G02500	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G16030.1)	4	-
187	249556_at	AT5G38195	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	4	-
188	261533_at	AT1G71690	similar to unknown protein [Arabidopsis thaliana]	4	-
189	249549_at	AT5G38180	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	4	-
190	248910_at	AT5G45820	CIPK20 (CBL-INTERACTING PROTEIN KINASE 20)	3	-
191	248764_at	AT5G47640	CCAAT-box binding transcription factor subunit B (NF-YB) (HAP3) (AHAP3) family (Hap3b)	3	-
192	261134_at	AT1G19630	CYP722A1 (cytochrome P450, family 722, subfamily A, polypeptide 1)	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

193	256453_at	AT1G75270	DHAR2; glutathione dehydrogenase (ascorbate)	3	-
194	253434_at	AT4G32500	AKT5 (Arabidopsis K+ transporter 5)	4	-
195	263258_at	AT1G10540	xanthine/uracil permease family protein	4	-
196	245832_at	AT1G48850	EMB1144 (EMBRYO DEFECTIVE 1144)	4	-
197	254307_at	AT4G22400	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G18320.1)	4	-
198	253579_at	AT4G30610	BRS1 (BRI1 SUPPRESSOR 1)	2	[Zhou and Li, 2005]
199	262238_at	AT1G48300	similar to hypothetical protein [Vitis vinifera] (GB:CAN81152.1)	4	-
200	247025_at	AT5G67030	ABA1 (ABA DEFICIENT 1)	2	[Barrero et al., 2008]; [Hemm et al., 2004]
201	262259_s_at	AT1G53870	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G53890.1)	4	-
202	257130_at	AT3G20210	DELTA-VPE (delta vacuolar processing enzyme)	2	[Nakaune et al., 2005]
203	248217_at	AT5G53560	ATB5-A (Cytochrome b5 A)	3	-
204	254957_at	AT4G10970	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G23910.1)	4	-
205	263010_at	AT1G23330	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G10740.1)	4	-
206	250633_at	AT5G07460	PMSR2 (PEPTIDEMETHIONINE SULFOXIDE REDUCTASE 2)	4	-
207	251520_at	AT3G59410	protein kinase family protein	3	-
208	248337_at	AT5G52310	COR78 (COLD REGULATED 78)	3	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

209	253433_s_at	AT4G28365	plastocyanin-like domain-containing protein	4	-
210	246180_at	AT5G20840	phosphoinositide phosphatase family protein	4	-
211	251497_at	AT3G59060	PIL6 (PHYTOCHROME-INTERACTING FACTOR 5)	2	[Hornitschek et al., 2012]
212	245488_at	AT4G16270	peroxidase 40 (PER40) (P40)	4	-
213	265605_at	AT2G25540	CESA10 (CELLULOSE SYNTHASE 10)	2	[Li et al., 2013]
214	252829_at	AT4G40060	ATHB-16/ATHB16 (ARABIDOPSIS THALIANA HOMEOBOX PROTEIN 16)	2	[Lechner et al., 2011]
215	250719_at	AT5G06250	transcription factor	2	[Causier et al., 2012]
216	251466_at	AT3G59340	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G59310.1)	4	-
217	251888_at	AT3G54190	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G38630.1)	4	-
218	253574_at	AT4G31030	contains domain PROKAR_LIPOPROTEIN (PS51257)	4	-
219	249477_s_at	AT5G38940	ion binding / metal ion binding / nutrient reservoir	4	-
220	250517_at	AT5G08260	SCPL35 (serine carboxypeptidase-like 35)	4	-
221	254777_at	AT4G12960	gamma interferon responsive lysosomal thiol reductase family protein / GILT family protein	3	-
222	251374_at	AT3G60390	HAT3 (homeobox-leucine zipper protein 3)	3	-
223	245999_at	AT5G20650	COPT5 (copper transporter 5)	3	-
224	264338_at	AT1G70300	KUP6 (K+ uptake permease 6)	3	-
225	248639_at	AT5G48930	transferase family protein	1	[Hoffmann et al., 2004]

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

226	258047_at	AT3G21240	4CL2 (4-coumarate:CoA ligase 2)	1	[Harding et al., 2002]
227	246466_at	AT5G17010	sugar transporter family protein	4	-
228	261925_at	AT1G22540	proton-dependent oligopeptide transport (POT) family protein	4	-
229	267240_at	AT2G02680	DC1 domain-containing protein	4	-
230	255554_at	AT4G01897	similar to unnamed protein product [Vitis vinifera] (GB:CAO40169.1)	4	-
231	266963_at	AT2G39450	ATMTP11/MTP11; cation transmembrane transporter/ manganese ion transmembrane transporter/ manganese:hydrogen antiporter	3	-
232	251735_at	AT3G56090	ATFER3 (FERRITIN 3)	2	[Tarantino et al., 2003]
233	259489_at	AT1G15790	similar to protein binding / transcription cofactor [Arabidopsis thaliana] (TAIR:AT1G15780.1)	4	-
234	262436_at	AT1G47610	transducin family protein / WD-40 repeat family protein	4	-
235	259853_at	AT1G72300	leucine-rich repeat transmembrane protein kinase, putative	2	[Amano et al., 2007]
236	248793_at	AT5G47240	ATNUDT8 (Arabidopsis thaliana Nudix hydrolase homolog 8)	3	-
237	253305_at	AT4G33666	unknown protein	4	-
238	264383_at	AT2G25080	ATGPX1 (GLUTATHIONE PEROXIDASE 1)	2	[Chang et al., 2009]
239	249198_s_at	AT5G42350	kelch repeat-containing F-box family protein	4	-
240	258116_at	AT3G14520	terpene synthase/cyclase family protein	4	-
241	247765_at	AT5G58860	CYP86A1 (cytochrome P450, family 86, subfamily A, polypeptide 1)	2	[Höfer et al., 2008]

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

242	247035_at	AT5G67110	ALC (ALCATRAZ)	2	[Groszmann et al., 2011]
243	247921_at	AT5G57660	zinc finger (B-box type) family protein	3	-
244	266110_at	AT2G02080	ATIDD4 (ARABIDOPSIS THALIANA INDETERMINATE(ID)-DOMAIN 4)	4	-
245	246419_at	AT5G17030	UDP-glucoronosyl/UDP-glucosyl transferase family protein	1	[Yonekura-Sakakibara et al., 2008]
246	266672_at	AT2G29650	inorganic phosphate transporter, putative	2	[Wang et al., 2011]
247	259292_at	AT3G11560	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G06220.1)	4	-
248	264066_at	AT2G27880	argonaute protein, putative / AGO, putative	3	-
249	245573_at	AT4G14730	Bax inhibitor-1 family protein	3	-
250	260124_at	AT1G36340	UBC31 (UBIQUITIN-CONJUGATING ENZYME 31)	4	-
251	248789_at	AT5G47440	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G17350.1)	4	-
252	266169_at	AT2G38900	serine protease inhibitor, potato inhibitor I-type family protein	4	-
253	255574_at	AT4G01420	CBL5 (CALCINEURIN B-LIKE PROTEIN 5)	3	-
254	261826_at	AT1G11580	ATPMEPCRA; pectinesterase	3	-
255	260024_at	AT1G30080	glycosyl hydrolase family 17 protein	4	-
256	245809_at	AT1G58440	XF1 (SQUALENE EPOXIDASE 1)	2	[Posé et al., 2009]
257	256985_at	AT3G13540	ATMYB5 (MYB DOMAIN PROTEIN 5)	2	[Li et al., 2009]
258	246374_at	AT1G51840	protein kinase-related	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

259	256137_at	AT1G48690	auxin-responsive GH3 family protein	4	-
260	261164_at	AT1G34470	permease-related	4	-
261	255954_at	AT1G22090	EMB2204 (EMBRYO DEFECTIVE 2204)	4	-
262	258130_at	AT3G24510	Encodes a defensin-like (DEFL) family protein	4	-
263	264026_at	AT2G21060	ATGRP2B (GLYCINE-RICH PROTEIN 2B)	3	-
264	250804_at	AT5G05030	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G11660.1)	4	-
265	260655_at	AT1G19320	pathogenesis-related thaumatin family protein	4	-
266	250467_at	AT5G10100	trehalose-6-phosphate phosphatase, putative	4	-
267	245389_at	AT4G17480	palmitoyl protein thioesterase family protein	4	-
268	264271_at	AT1G60270	pseudogene, glycosyl hydrolase family 1	4	-
269	261855_at	AT1G50510	indigoidine synthase A-like protein	4	-
270	253769_at	AT4G28560	RIC7 (ROP-INTERACTIVE CRIB MOTIF-CONTAINING PROTEIN 7)	4	-
271	263995_at	AT2G22540	MADS-box protein SVP	2	[Seo et al., 2009]
272	245020_at	ATCG00540	Encodes cytochrome f apoprotein	4	-
273	254693_at	AT4G17880	basic helix-loop-helix (bHLH) family protein	2	[Fernández-Calvo et al., 2011]
274	256948_at	AT3G18930	zinc finger (C3HC4-type RING finger) family protein	4	-
275	255432_at	AT4G03330	SYP123 (syntaxin 123); SNAP receptor	3	-
276	245929_at	AT5G24760	alcohol dehydrogenase, putative	4	-
277	247113_at	AT5G65960	similar to CM0216.310.nc [Lotus japonicus] (GB:BAF98215.1)	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

278	249259_at	AT5G41660	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G44430.1)	4	-
279	260806_at	AT1G78260	RNA recognition motif (RRM)-containing protein	4	-
280	253582_at	AT4G30670	contains domain PROKAR_LIPOPROTEIN (PS51257)	4	-
281	264853_at	AT2G17260	GLR2 (GLUTAMATE RECEPTOR 2)	3	-
282	266086_at	AT2G38060	transporter-related	4	-
283	258903_at	AT3G06410	nucleic acid binding	4	-
284	267226_at	AT2G44010	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G59880.1)	4	-
285	265053_at	AT1G52000	jacalin lectin family protein	4	-
286	247468_at	AT5G62000	ARF2 (AUXIN RESPONSE FACTOR 2)	2	[Smaczniak et al., 2012]
287	251022_at	AT5G02150	binding	4	-
288	267394_s_at	AT2G44550	ATGH9B10 (ARABIDOPSIS THALIANA GLYCOSYL HYDROLASE 9B10)	4	-
289	267122_at	AT2G23550	hydrolase	4	-
290	255730_at	AT1G25460	oxidoreductase family protein	4	-
291	267552_at	AT2G32770	ATPAP13/PAP13; acid phosphatase	3	-
292	262103_at	AT1G02940	ATGSTF5 (Arabidopsis thaliana Glutathione S-transferase (class phi) 5)	3	-
293	264643_at	AT1G08990	PGSIP5 (PLANT GLYCOGENIN-LIKE STARCH INITIATION PROTEIN 5)	4	-
294	253166_at	AT4G35290	GLUR2 (Glutamate receptor 2)	3	-
295	251295_at	AT3G62000	O-methyltransferase family 3 protein	4	-
296	247795_at	AT5G58620	zinc finger (CCCH-type) family protein	3	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

297	248202_at	AT5G54220	Encodes a defensin-like (DEFL) family protein	4	-
298	263027_at	AT1G24010	Identical to Uncharacterized protein At1g24010 [Arabidopsis Thaliana] (GB:P0C0B1;GB:Q9LR93)	4	-
299	247796_at	AT5G58782	dehydrodolichyl diphosphate synthase, putative / DEDOL-PP synthase, putative	3	-
300	250360_at	AT5G11360	ATP binding / protein kinase	4	-
301	250168_at	AT5G15320	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G01130.1)	4	-
302	251674_at	AT3G57250	emsy N terminus domain-containing protein / ENT domain-containing protein	4	-
303	257039_at	AT3G19160	ATIPT8 (Arabidopsis thaliana isopentenyltransferase 8); adenylate dimethylallyltransferase	2	[Takei et al., 2004]
304	264297_at	AT1G78710	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G31110.2)	4	-
305	265986_at	AT2G24230	leucine-rich repeat transmembrane protein kinase, putative	4	-
306	255028_at	AT4G09890	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G47480.1)	4	-
307	260153_at	AT1G52760	esterase/lipase/thioesterase family protein	3	-
308	250701_at	AT5G06839	bZIP family transcription factor	2	[Murmu et al., 2010]
309	261848_at	AT1G11590	pectin methylesterase, putative	4	-
310	248652_at	AT5G49270	COBL9/MRH4/SHV2 (COBRA-LIKE 9, SHAVEN 2)	3	-
311	246887_at	AT5G26250	sugar transporter, putative	3	-
312	267256_s_at	AT2G23000	SCPL10 (serine carboxypeptidase-like 10)	2	[Fraser et al., 2007]

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

313	250477_at	AT5G10190	transporter-related	4	-
314	266257_at	AT2G27820	PD1 (PREPHENATE DEHYDRATASE 1)	2	[Cho et al., 2007]
315	250381_at	AT5G11610	exostosin family protein	4	-
316	245304_at	AT4G15630	integral membrane family protein	3	-
317	259568_at	AT1G20490	AMP-dependent synthetase and ligase family protein	4	-
318	254682_at	AT4G13640	UNE16 (unfertilized embryo sac 16)	2	[Klopffleisch et al., 2011]
319	260630_at	AT1G62340	ALE1 (ABNORMAL LEAF SHAPE 1)	3	-
320	248761_at	AT5G47635	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G40113.1)	4	-
321	250234_at	AT5G13420	transaldolase, putative	4	-
322	246919_at	AT5G25460	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G11420.1)	4	-
323	248042_at	AT5G55960	similar to unnamed protein product [Vitis vinifera] (GB:CAO45175.1)	4	-
324	246716_s_at	AT5G28960	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT5G28910.2)	4	-
325	254906_at	AT4G11180	disease resistance-responsive family protein / dirigent family protein	4	-
326	259567_at	AT1G20500	4-coumarate–CoA ligase-like 4	4	-
327	248154_at	AT5G54400	methyltransferase	4	-
328	254900_at	AT4G11510	RALFL28 (RALF-LIKE 28)	4	-
329	252342_at	AT3G48950	glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein	4	-
330	247383_at	AT5G63410	leucine-rich repeat transmembrane protein kinase, putative	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

331	251984_at	AT3G53260	PAL2 (phenylalanine ammonia-lyase 2)	1	[Olsen et al., 2008]
332	266234_at	AT2G02350	SKIP3 (SKP1 INTERACTING PARTNER 3)	3	-
333	257654_at	AT3G13310	DNAJ heat shock N-terminal domain-containing protein	3	-
334	266808_at	AT2G29995	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G07175.1)	4	-
335	263177_at	AT1G05540	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G30160.2)	4	-
336	247416_at	AT5G63070	40S ribosomal protein S15, putative	4	-
337	249491_at	AT5G39130	germin-like protein, putative	4	-
338	255294_at	AT4G04750	carbohydrate transmembrane transporter/ sugar:hydrogen ion symporter	4	-
339	264010_at	AT2G21100	disease resistance-responsive protein-related / dirigent protein-related	4	-
340	255687_at	AT4G00640	unknown protein	4	-
341	245097_at	AT2G40935	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G18470.1)	4	-
342	256872_at	AT3G26490	phototropic-responsive NPH3 family protein	4	-
343	261487_at	AT1G14340	RNA recognition motif (RRM)-containing protein	4	-
344	255025_at	AT4G09900	hydrolase, alpha/beta fold family protein	4	-
345	262939_s_at	AT1G79530	GAPCP-1; glyceraldehyde-3-phosphate dehydrogenase	3	-
346	263128_at	AT1G78600	zinc finger (B-box type) family protein	1	[Datta et al., 2008]
347	258342_at	AT3G22800	leucine-rich repeat family protein / extensin family protein	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

348	251133_at	AT5G01240	amino acid permease, putative	2	[Yang et al., 2012]
349	252365_at	AT3G48350	cysteine proteinase, putative	3	-
350	256328_at	AT3G02360	6-phosphogluconate dehydrogenase family protein	2	[Swatek et al., 2011]
351	256491_at	AT1G31500	endonuclease/exonuclease/phosphatase family protein	4	-
352	266311_at	AT2G27130	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	4	-
353	247110_at	AT5G65830	leucine-rich repeat family protein	4	-
354	254202_at	AT4G24140	hydrolase, alpha/beta fold family protein	4	-
355	252831_at	AT4G39980	DHS1 (3-DEOXY-D-ARABINO-HEPTULOSONATE 7-PHOSPHATE SYNTHASE 1)	4	-
356	253394_at	AT4G32770	VTE1 (VITAMIN E DEFICIENT 1)	2	[Semchuk et al., 2009]
357	255345_at	AT4G04460	aspartyl protease family protein	4	-
358	253463_at	AT4G32105	galactosyltransferase	4	-
359	261247_at	AT1G20070	unknown protein	4	-
360	248371_at	AT5G51810	AT2353/ATGA20OX2/GA20OX2 (GIBBERELLIN 20 OXIDASE 2)	2	[Rieu et al., 2007]
361	264504_at	AT1G09430	ACLA-3 (ATP-citrate lyase A-3)	2	[Fatland et al., 2005]; [Fatland et al., 2002]
362	261907_at	AT1G65060	4CL3 (4-coumarate:CoA ligase 3)	1	[Ehlting et al., 2002]
363	265645_at	AT2G27370	integral membrane family protein	4	-
364	247713_at	AT5G59330	Encodes a Protease inhibitor/seed storage/LTP family protein [pseudogene]	4	-

Table 5.1: Description of 382 expansion genes of the flavonoids pathway.

365	259598_at	AT1G27980	pyridoxal-dependent decarboxylase family protein	2	[Nishikawa et al., 2008]
366	251448_at	AT3G59845	NADP-dependent oxidoreductase, putative	4	-
367	248795_at	AT5G47390	myb family transcription factor	4	-
368	255798_at	AT2G33255	hydrolase	4	-
369	245130_at	AT2G45340	leucine-rich repeat transmembrane protein kinase, putative	4	-
370	250473_at	AT5G10220	ANN6 (ANNEXIN ARABIDOPSIS 6)	3	-
371	258960_at	AT3G10590	myb family transcription factor	4	-
372	267125_at	AT2G23580	hydrolase, alpha/beta fold family protein	3	-
373	263114_at	AT1G03130	PSAD-2 (photosystem I subunit D-2)	2	[Yu et al., 2008]
374	253886_at	AT4G27710	CYP709B3 (cytochrome P450, family 709, subfamily B, polypeptide 3)	2	[Huang et al., 2006]
375	251750_at	AT3G55710	UDP-glucoronosyl/UDP-glucosyl transferase family protein	4	-
376	259129_at	AT3G02150	PTF1 (PLASTID TRANSCRIPTION FACTOR 1)	2	[Baba et al., 2001]
377	250909_at	AT5G03700	PAN domain-containing protein	4	-
378	259441_at	AT1G02300	cathepsin B-like cysteine protease, putative	4	-
379	266578_at	AT2G23910	cinnamoyl-CoA reductase-related	4	-
380	264352_at	AT1G03270	similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G14240.1)	4	-
381	265511_at	AT2G05540	glycine-rich protein	4	-
382	263203_at	AT1G05490	CHR31 (chromatin remodeling 31)	3	-

Table 5.4: **Description of 382 expansion genes of the flavonoids pathway.** [rnk = ranking]

Chapter 6

Conclusions

In recent years, algorithms used to expand gene regulatory network (Section 2.3) appeared together with algorithms used to infere gene regulatory network (Section 2.1). The expansion can be done at two levels. At the first level, object of the expansion is to find new genes and the relationships between these genes and the genes of the gene network. At the second level, the aim is simpler, that is identifying new genes which expand the known network without taking care of the relationships.

PC-IM, the method proposed in this thesis, has been developed for this second purpose. Its main characteristics are:

- use of *a priori* information about a known gene regulatory network (Local Gene Network (LGN));
- possibility to consider all the genes of the input dataset (e.g. all genes of a genome or all genes (or probes) of a gene expression experiment);
- use of observational gene expression data;
- intrinsic capacity to estimate its performances;
- iteration, namely each whole procedure of LGN expansion is repeated a number of times i;
- possibility to exchange the PC-algorithm with another algorithm.

PC-IM can deal with a large number of genes (all the genes of the input dataset), because it divides the input gene list in *tiles* of the same size. The genes of the LGN (intra genes) are present in each *tile*, while the genes of the input list (extra genes) are present only in one *tile* for each iteration. After the run, the PC-algorithm output is a list of extra genes and intra genes for each single iteration. A frequency value is also given for

each single gene of the list. All the lists are then combinated and a normalized frequency value is calculated.

The intrinsic evaluation of the performances is obtained by estimation of PPV, Se and 1-Sp paramteters. Values of these parameters are obtained dividing the LGN in different subLGNs and estimating precision, sensitivity ans specificity of the expansion between the genes of each subLGNs and the other genes of the LGN. The prior knowledge about LGN is used to calculate the number of TP, FP, FN and TN. Finally, PC-IM, finds the cut-off frequency value that gives an expansion genes list with maximum performances (Chapter 3). This cut-off frequency is applyed on the final output of PC-IM. The final output is a list of extra genes related with the genes of the LGN.

In Chapter 4 preliminary evaluation tests (Section 4.1 and Section 4.1.2) and PC-IM evaluation tests (Section 4.3) were performed.

The aims of the preliminary evaluation tests were:

- evaluate the influence of the type of gene expression data (*in silico* or *in vivo* gene expression data);

- judge the opportunity to use the PC algorithm in running PC-IM.

The results showed that the values of precision and sensitivity change appreciably between *in silico* and *in vivo* data and between different *in vivo* data. These changes underline the importance of the choice of the type of gene expression data and how the generation of the *in silico* data can influence the performance of algorithms. Our choice, to use *in vivo* gene expression data, derives from this consideration and from the fact that in public databases there is a large availability of observational gene expression data.

The aims of the evaluation tests (Section 4.3) were:

- selection of PC-IM parameters (*tile* size), number of iterations, type of gene expression data that give the best intrinsic performances;

- understand how the topology of the LGN influences the performances of PC-IM (Section 4.3.4);

- comparison of PC-IM with another recently proposed expansion method (GENIES) (Section 4.3.6).

The evaluation of the *tile* size showed that 100 and 1000 is the number of genes that gives the best instrinsic performance with PC. The size of the *tiles* clearly depends on the algorithm used in PC-IM. Our result with PC confirmed the results obtained by Wang et al. [2010].

The iteration evaluation showed that the best performances are reached when the iteration numer is equal or greater than 50.

The evaluation of PC-IM highlighted also the importance of the data. A mix between expression data related with the metabolism is part of the LGN and gene expression data related to it.

The study, on the effect of the type of LGN, indicated that PC-IM is robust on this regard. In fact PC-IM provides significantly different results when as input is given a real LGN or a random LGN (Section 4.3.4).

The comparison between PC-IM and GENIES showed that GENIES gives the best expansion performances of the subLGNs, but it does not find extra genes. Moreover in GENIES, to get to the best performances is necessary to test different combinations between kernel matrix and algorithms. From these considerations was evident that GENIES is not suitable tool to expand a LGN with extra genes and that GENIES is not easy to use for a user without informatic knowledge.

Chapter 5 reports applications of PC-IM to expand two different LGNs (Section 5.1 and Section 5.2) of *Arabidopsis thaliana*.

The validation of the expansion genes, given as output by PC-IM and selected by cut-off frequency value, was done by bibliographic search. In particular it was seen that the top ranking expansion genes were very correlated in genes related to the metabolism the LGN was part of. The final proof of the proposed expansion list will come from *in vivo* experiments, which shall confirm or not whether the new genes are functionally related to the starting LGN.

Dissemination of results

The results of this thesis work have been disseminated by:

- Poster presentation:
  Coller E., Malacarne G., Moser C., Blanzieri E., Application of the PC algorithm to infer regulatory networks from observational gene expression data. *CMSB2010*, September 29-October 1, 2010, Trento, Italy.

- Patent application:
  European Patent application EP13151728.6 (of date 17 January 2013)
  Title: SYSTEMS AND METHODS FOR DETERMINING SUITABLE ENTITIES FOR EXPANDING ESTABLISHED CAUSAL MOLECULAR BIOLOGICAL NETWORKS AND FOR DETERMINING SIGNIFICANT CAUSAL RELATIONSHIPS BETWEEN ENTITIES OF ESTABLISHED CAUSAL MOLECULAR BIOLOGICAL NETWORKS AND CANDIDATE ENTITIES.
  Inventors: Enrico Blanzieri, **Emanuela Coller**, Giulia Malacarne, Claudio Moser
  Applicants: Fondazione Edmund Mach (San Michele all'Adige, Italy) and Universitá degli Studi di Trento.

- Journal article (in preparation)

Bibliography

Abbott, Derek A; Suir, Erwin; van Maris, Antonius JA, and Pronk, Jack T. Physiological and transcriptional responses to high concentrations of lactic acid in anaerobic chemostat cultures of *Saccharomyces cerevisiae*. *Applied and Environmental Microbiology*, 74(18):5759–5768, 2008.

Agarwal, Ameeta K; Xu, Tao; Jacob, Melissa R; Feng, Qin; Lorenz, Michael C; Walker, Larry A, and Clark, Alice M. Role of heme in the antifungal activity of the azaoxoaporphine alkaloid sampangine. *Eukaryotic Cell*, 7(2):387–400, 2008.

Aguilera, Jaime; Petit, Thomas; Winde, Johannes H, and Pronk, Jack T. Physiological and genome-wide transcriptional responses of saccharomyces cerevisiae to high carbon dioxide concentrations. *FEMS Yeast Research*, 5(6-7):579–593, 2006.

Agullo, Georgine; Gamet-Payrastre, Laurence; Manenti, Stéphane; Viala, Cécile; Rémésy, Christian; Chap, Hugues, and Payrastre, Bernard. Relationship between flavonoid structure and inhibition of phosphatidylinositol 3-kinase: a comparison with tyrosine kinase and protein kinase c inhibition. *Biochemical Pharmacology*, 53(11):1649–1657, 1997.

Ajjawi, Imad; Rodriguez Milla, Miguel A; Cushman, John, and Shintani, David K. Thiamin pyrophosphokinase is required for thiamin cofactor activation in arabidopsis. *Plant Molecular Biology*, 65(1):151–162, 2007.

Altay, Gökmen and Emmert-Streib, Frank. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, 26(14):1738–1744, 2010.

Aluri, Sirisha and Büttner, Michael. Identification and functional expression of the *Arabidopsis thaliana* vacuolar glucose transporter 1 and its role in seed germination and flowering. *Proceedings of the National Academy of Sciences*, 104(7):2537–2542, 2007.

Alves-Ferreira, Márcio; Wellmer, Frank; Banhara, Aline; Kumar, Vijaya; Riechmann, José Luis, and Meyerowitz, Elliot M. Global expression profiling applied to the analysis of arabidopsis stamen development. *Plant Physiology*, 145(3):747–762, 2007.

Amano, Yukari; Tsubouchi, Hiroko; Shinohara, Hidefumi; Ogawa, Mari, and Matsubayashi, Yoshikatsu. Tyrosine-sulfated glycopeptide involved in cellular proliferation and expansion in arabidopsis. *Proceedings of the National Academy of Sciences*, 104(46):18333–18338, 2007.

Antonakis, John; Bendahan, Samuel; Jacquart, Philippe, and Lalive, Rafael. On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6):1086–1120, 2010.

Aragon, Anthony D; Rodriguez, Angelina L; Meirelles, Osorio; Roy, Sushmita; Davidson, George S; Tapia, Phillip H; Allen, Chris; Joe, Ray; Benn, Don, and Werner-Washburne, Margaret. Characterization of differentiated quiescent and nonquiescent cells in yeast stationary-phase cultures. *Molecular Biology of the Cell*, 19(3):1271–1280, 2008.

Auld, Kathryn L; Brown, Christopher R; Casolari, Jason M; Komili, Suzanne, and Silver, Pamela A. Genomic association of the proteasome demonstrates overlapping gene regulatory activity with transcription factor substrates. *Molecular Cell*, 21(6):861–871, 2006.

Ausin, Israel; Mockler, Todd C; Chory, Joanne, and Jacobsen, Steven E. IDN1 and IDN2 are required for de novo DNA methylation in *Arabidopsis thaliana*. *Nature Structural & Molecular Biology*, 16(12):1325–1327, 2009.

Azzouz, Nowel; Panasenko, Olesya O; Deluen, Cécile; Hsieh, Julien; Theiler, Grégory, and Collart, Martine A. Specific roles for the ccr4-not complex subunits in expression of the genome. Rna, 15(3):377–383, 2009.

Baba, Kyoko; Nakano, Takeshi; Yamagishi, Kazutoshi, and Yoshida, Shigeo. Involvement of a nuclear-encoded basic helix-loop-helix protein in transcription of the light-responsive promoter ofpsbD. *Plant Physiology*, 125 (2):595–603, 2001.

Bach, Liên and Faure, Jean-Denis. Role of very-long-chain fatty acids in plant development, when chain length does matter. *Comptes Rendus Biologies*, 333(4):361–370, 2010.

Bansal, Mukesh; Della Gatta, Giusy, and Di Bernardo, Diego. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

Bansal, Mukesh; Belcastro, Vincenzo; Ambesi-Impiombato, Alberto, and Di Bernardo, Diego. How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3(1), 2007.

Barbara, Kellie E; Haley, Terry M; Willis, Kristine A, and Santangelo, George M. The transcription factor gcr1 stimulates cell growth by participating in nutrient-responsive gene expression on a global level. *Molecular Genetics and Genomics*, 277(2):171–188, 2007.

Barrero, JosÉ; Rodríguez, Pedro L; Quesada, VICtor; AlabadI, David; Blázquez, Miguel A; Boutin, Jean-Pierre; Marion-Poll, Annie; Ponce, María Rosa, and Micol, José Luis. The ABA1 gene and carotenoid biosynthesis are required for late skotomorphogenic growth in *Arabidopsis thaliana*. *Plant Cell & Environment*, 31(2):227–234, 2008.

Bartel, Paul L and Fields, Stanley. *The yeast two-hybrid system*. Oxford University Press, USA, 1997.

Basso, Katia; Margolin, Adam A; Stolovitzky, Gustavo; Klein, Ulf; Dalla-Favera, Riccardo, and Califano, Andrea. Reverse engineering of regulatory networks in human b cells. *Nature Genetics*, 37(4):382–390, 2005.

Bektas, Inga; Fellenberg, Christin, and Paulsen, Harald. Water-soluble chlorophyll protein (WSCP) of *Arabidopsis* is expressed in the gynoecium and developing silique. *Planta*, pages 1–9, 2012.

Bemer, Marian; Wolters-Arts, Mieke; Grossniklaus, Ueli, and Angenent, Gerco C. The MADS domain protein DIANA acts together with AGAMOUS-LIKE80 to specify the central cell in arabidopsis ovules. *The Plant Cell Online*, 20(8):2088–2101, 2008.

Bennett, Simon T; Barnes, Colin; Cox, Anthony; Davies, Lisa, and Brown, Clive. Toward the 1000 human genome. *Pharmacogenomics*, 6(4):373–382, 2005.

Benton, Michael G; Somasundaram, Swetha; Glasner, Jeremy D, and Palecek, Sean P. Analyzing the dose-dependence of the *Saccharomyces cerevisiae* global transcriptional response to methyl methanesulfonate and ionizing radiation. *BMC Genomics*, 7(1):305, 2006.

Bernstein, Bradley E; Tong, Jeffrey K, and Schreiber, Stuart L. Genomewide studies of histone deacetylase function in yeast. *Proceedings of the National Academy of Sciences*, 97(25):13708–13713, 2000.

Berr, Alexandre; Xu, Lin; Gao, Juan; Cognat, Valérie; Steinmetz, Andre; Dong, Aiwu, and Shen, Wen-Hui. SET DOMAIN GROUP25 encodes a histone methyltransferase and is involved in FLOWERING LOCUS C activation and repression of flowering. *Plant Physiology*, 151(3):1476–1485, 2009.

Berry, David B and Gasch, Audrey P. Stress-activated genomic expression changes serve a preparative role for impending stress in yeast. *Molecular Biology of the Cell*, 19(11):4580–4587, 2008.

Besseau, Sébastien; Hoffmann, Laurent; Geoffroy, Pierrette; Lapierre, Catherine; Pollet, Brigitte, and Legrand, Michel. Flavonoid accumulation in arabidopsis repressed in lignin synthesis affects auxin transport and plant growth. *The Plant Cell Online*, 19(1):148–162, 2007.

Bies-Etheve, Natacha; Gaubier-Comella, Pascale; Debures, Anne; Lasserre, Eric; Jobet, Edouard; Raynal, Monique; Cooke, Richard, and Delseny, Michel. Inventory, evolution and expression profiling diversity of the LEA (late embryogenesis abundant) protein gene family in *Arabidopsis thaliana*. *Plant Molecular Biology*, 67(1):107–124, 2008.

Bolouri-Moghaddam, Mohammad Reza; Le Roy, Katrien; Xiang, Li; Rolland, Filip, and Van den Ende, Wim. Sugar signalling and antioxidant network connections in plant cells. *FEBS Journal*, 277(9):2022–2037, 2010.

Broun, Pierre. Transcriptional control of flavonoid biosynthesis: a complex network of conserved regulators involved in multiple aspects of differentiation in *Arabidopsis*. *Current Opinion in Plant Biology*, 8(3):272–279, 2005.

Cai, Suqin and Lashbrook, Coralie C. Stamen abscission zone transcriptome profiling reveals new candidates for abscission control: enhanced retention of floral organs in transgenic plants overexpressing arabidopsis ZINC FINGER PROTEIN2. *Plant Physiology*, 146(3):1305–1321, 2008.

Cai, Xiaoning; Ballif, Jenny; Endo, Saori; Davis, Elizabeth; Liang, Mingxiang; Chen, Dong; DeWald, Daryll; Kreps, Joel; Zhu, Tong, and Wu, Yajun. A putative CCAAT-binding transcription factor is a regulator of flowering timing in arabidopsis. *Plant Physiology*, 145(1):98–105, 2007.

Causier, Barry; Schwarz-Sommer, Zsuzsanna, and Davies, Brendan. Floral organ identity: 20 years of ABCs. *Seminars in Cell & Developmental Biology*, 21(1):73–79, 2010.

Causier, Barry; Ashworth, Mary; Guo, Wenjia, and Davies, Brendan. The TOPLESS interactome: a framework for gene repression in arabidopsis. *Plant Physiology*, 158(1):423–438, 2012.

Chang, Christine CC; Slesak, Ireneusz; Jordá, Lucía; Sotnikov, Alexey; Melzer, Michael; Miszalski, Zbigniew; Mullineaux, Philip M; Parker, Jane E; Karpinska, Barbara, and Karpinski, Stanislaw. Arabidopsis chloroplastic glutathione peroxidases play a role in cross talk between photooxidative stress and immune responses. *Plant Physiology*, 150(2):670–683, 2009.

Chee, Mark; Yang, Robert; Hubbell, Earl; Berno, Anthony; Huang, Xiaohua C; Stern, David; Winkler, Jim; Lockhart, David J; Morris, Macdonald S; Fodor, Stephen PA, and others, . Accessing genetic information with high-density dna arrays. *Science*, 274(5287):610–614, 1996.

Chen, Feng; Ro, Dae-Kyun; Petri, Jana; Gershenzon, Jonathan; Bohlmann, Jörg; Pichersky, Eran, and Tholl, Dorothea. Characterization of a root-specific arabidopsis terpene synthase responsible for the formation of the volatile monoterpene 1, 8-cineole. *Plant Physiology*, 135(4):1956–1966, 2004.

Chen, Shuo; Xing, Xin-Hui; Huang, Jian-Jun, and Xu, Ming-Shu. Enzyme-assisted extraction of flavonoids from *Ginkgo biloba* leaves: Improvement effect of flavonol transglycosylation catalyzed by *Penicillium decumbens* cellulase. *Enzyme and Microbial Technology*, 48(1):100–105, 2011.

Chen, Xiaobo; Zhang, Zenglin; Liu, Danmei; Zhang, Kai; Li, Aili, and Mao, Long. SQUAMOSA promoter-binding protein-like transcription factors: Star players for plant growth and development. *Journal of Integrative Plant Biology*, 52(11):946–951, 2010.

Chen, Yanhui; Zhang, Xiangbo; Wu, Wei; Chen, Zhangliang; Gu, Hongya, and Qu, Li-Jia. Overexpression of the wounding-responsive gene atMYB15 activates the shikimate pathway in arabidopsis. *Journal of Integrative Plant Biology*, 48(9):1084–1095, 2006.

Cho, Man-Ho; Corea, Oliver RA; Yang, Hong; Bedgar, Diana L; Laskar, Dhrubojyoti D; Anterola, Aldwin M; Moog-Anterola, Frances Anne; Hood, Rebecca L; Kohalmi, Susanne E; Bernards, Mark A, and others, . Phenylalanine biosynthesis in *Arabidopsis thaliana*. *Journal of Biological Chemistry*, 282(42):30827–30835, 2007.

Chu, Tianjiao; Glymour, Clark; Scheines, Richard, and Spirtes, Peter. A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19 (9):1147–1152, 2003.

Clark, Steven E; Running, Mark P, and Meyerowitz, Elliot M. CLAVATA1, a regulator of meristem and flower development in arabidopsis. *Development*, 119(2):397–418, 1993.

Coen, Enrico S; Meyerowitz, Elliot M, and others, . The war of the whorls: genetic interactions controlling flower development. *Nature*, 353(6339):31–37, 1991.

Corbesier, Laurent; Vincent, Coral; Jang, Seonghoe; Fornara, Fabio; Fan, Qingzhi; Searle, Iain; Giakountis, Antonis; Farrona, Sara; Gissot, Lionel; Turnbull, Colin, and others, . Ft protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. *Science Signalling*, 316(5827):1030, 2007.

Cosio, Claudia and Dunand, Christophe. Transcriptome analysis of various flower and silique development stages indicates a set of class iii peroxidase genes potentially involved in pod shattering in *Arabidopsis thaliana*. *BMC Genomics*, 11(1):528, 2010.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. Wiley-interscience, 2006.

D Auria, John C; Reichelt, Michael; Luck, Katrin; Svatoš, Aleš, and Gershenzon, Jonathan. Identification and characterization of the bahd acyltransferase malonyl coa: Anthocyanidin 5-o-glucoside-6-o-malonyltransferase (at5MAT) in *Arabidopsis thaliana*. FEBS *letters*, 581(5):872–878, 2007.

Daran-Lapujade, Pascale; Rossell, Sergio; Van Gulik, Walter M; Luttik, Marijke AH; de Groot, Marco JL; Slijper, Monique; Heck, Albert JR; Daran, Jean-Marc; de Winde, Johannes H; Westerhoff, Hans V, and others, . The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proceedings of the National Academy of Sciences*, 104(40):15753–15758, 2007.

Dardick, Christopher D; Callahan, Ann M; Chiozzotto, Remo; Schaffer, Robert J; Piagnani, M Claudia, and Scorza, Ralph. Stone formation in peach fruit exhibits spatial coordination of the lignin and flavonoid pathways and similarity to arabidopsis dehiscence. *BMC biology*, 8(1):13, 2010.

Das, Partha M; Ramachandran, Kavitha; vanWert, Jane; Singal, Rakesh, and others, . Chromatin immunoprecipitation assay. *Biotechniques*, 37(6):961, 2004.

Datta, Sourav; Johansson, Henrik; Hettiarachchi, Chamari; Irigoyen, María Luisa; Desai, Mintu; Rubio, Vicente, and Holm, Magnus. LZF1/SALT TOLERANCE HOMOLOG3, an arabidopsis B-box protein involved in light-dependent development and gene expression, undergoes COP1-mediated ubiquitination. *The Plant Cell Online*, 20(9):2324–2338, 2008.

Davidson, Eric H; Rast, Jonathan P; Oliveri, Paola; Ransick, Andrew; Calestani, Cristina; Yuh, Chiou-Hwa; Minokawa, Takuya; Amore, Gabriele; Hinman, Veronica; Arenas-Mena, Cesar, and others, . A genomic regulatory network for development. *Science Signalling*, 295(5560):1669, 2002.

Deikman, Jill and Hammer, Philip E. Induction of anthocyanin accumulation by cytokinins in *Arabidopsis thaliana*. *Plant Physiology*, 108(1):47–57, 1995.

Deng, Weiwei; Ying, Hua; Helliwell, Chris A; Taylor, Jennifer M; Peacock, W James, and Dennis, Elizabeth S. FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of arabidopsis. *Proceedings of the National Academy of Sciences*, 108(16):6680–6685, 2011.

di Bernardo, Diego; Thompson, Michael J; Gardner, Timothy S; Chobot, Sarah E; Eastwood, Erin L; Wojtovich, Andrew P; Elliott, Sean J; Schaus, Scott E, and Collins, James J. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology*, 23(3):377–383, 2005.

Dinkins, Randy D; Tavva, Venkata S; Palli, S Reddy, and Collins, Glenn B. Mutant and overexpression analysis of a C2H2 single zinc finger gene of arabidopsis. *Plant Molecular Biology Reporter*, 30(1):99–110, 2012.

Ditta, Gary; Pinyopich, Anusak; Robles, Pedro; Pelaz, Soraya, and Yanofsky, Martin F. The SEP4 gene of *Arabidopsis thaliana* functions in floral organ and meristem identity. *Current Biology*, 14(21):1935–1940, 2004.

Do, Cao-Trung; Pollet, Brigitte; Thévenin, Johanne; Sibout, Richard; Denoue, Dominique; Barrière, Yves; Lapierre, Catherine, and Jouanin, Lise. Both caffeoyl coenzyme a 3-o-methyltransferase 1 and caffeic acid o-methyltransferase 1 are involved in redundant functions for lignin, flavonoids and sinapoyl malate biosynthesis in arabidopsis. *Planta*, 226(5):1117–1129, 2007.

Dorca-Fornell, Carmen; Gregis, Veronica; Grandi, Valentina; Coupland, George; Colombo, Lucia, and Kater, Martin M. The arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems. *The Plant Journal*, 67(6):1006–1017, 2011.

Dou, Xiao-Ying; Yang, Ke-Zhen; Zhang, Yi; Wang, Wei; Liu, Xiao-Lei; Chen, Li-Qun; Zhang, Xue-Qin, and Ye, De. WBC27, an adenosine tri-phosphate-binding cassette protein, controls pollen wall formation and patterning in arabidopsis. *Journal of Integrative Plant Biology*, 53(1):74–88, 2011.

Ebrahimi, Ali; Aghdam, Rosa; Niloofar, Parisa; Ganjali, Mojtaba, and Eslahchi, Changiz. LSPC: An algorithm for inference of gene networks using bayesian network. *Journal of Emerging Trends in Computing and Information Sciences*, 3(5), 2012.

Ehlting, Jürgen; Büttner, Daniela; Wang, Qing; Douglas, Carl J; Somssich, Imre E, and Kombrink, Erich. Three 4-coumarate: coenzyme a ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in angiosperms. *The Plant Journal*, 19(1):9–20, 2002.

Eisen, Michael B; Spellman, Paul T; Brown, Patrick O, and Botstein, David. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

Emmert-Streib, Frank; Glazko, Galina V; Altay, Gökmen, and de Matos Simoes, Ricardo. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in Genetics*, 3, 2012.

Espinosa-Soto, Carlos; Padilla-Longoria, Pablo, and Alvarez-Buylla, Elena R. A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *The Plant Cell Online*, 16(11):2923–2939, 2004.

Faith, Jeremiah J; Driscoll, Michael E; Fusaro, Vincent A; Cosgrove, Elissa J; Hayete, Boris; Juhn, Frank S; Schneider, Stephen J, and Gardner, Timothy S. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, 36(suppl 1):D866–D870, 2008.

Fatland, Beth L; Ke, Jinshan; Anderson, Marc D; Mentzen, Wieslawa I; Cui, Li Wei; Allred, C Christy; Johnston, Jerry L; Nikolau, Basil J, and Wurtele, Eve Syrkin. Molecular characterization of a heteromeric ATP-citrate lyase that generates cytosolic acetyl-coenzyme A in arabidopsis. *Plant Physiology*, 130(2):740–756, 2002.

Fatland, Beth L; Nikolau, Basil J, and Wurtele, Eve Syrkin. Reverse genetic characterization of cytosolic acetyl-coa generation by ATP-citrate lyase in arabidopsis. *The Plant Cell Online*, 17(1):182–203, 2005.

Favaro, Rebecca; Pinyopich, Anusak; Battaglia, Raffaella; Kooiker, Maarten; Borghi, Lorenzo; Ditta, Gary; Yanofsky, Martin F; Kater, Martin M, and Colombo, Lucia. MADS-box protein complexes control carpel and ovule development in arabidopsis. *The Plant Cell Online*, 15(11):2603–2611, 2003.

Fernandes, Andrew D and Gloor, Gregory B. Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics*, 26(9):1135–1139, 2010.

Fernández-Calvo, Patricia; Chini, Andrea; Fernández-Barbero, Gemma; Chico, José-Manuel; Gimenez-Ibanez, Selena; Geerinck, Jan; Eeckhout, Dominique; Schweizer, Fabian; Godoy, Marta; Franco-Zorrilla, José Manuel, and others, . The arabidopsis bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. *The Plant Cell Online*, 23(2):701–715, 2011.

Floerl, Saskia; Majcherczyk, Andrzej; Possienke, Mareike; Feussner, Kirstin; Tappe, Hella; Gatz, Christiane; Feussner, Ivo; Kües, Ursula, and Polle, Andrea. Verticillium longisporum infection affects the leaf apoplastic proteome, metabolome, and cell wall properties in *Arabidopsis thaliana*. *PloS One*, 7(2):e31435, 2012.

Fornara, Fabio and Coupland, George. Plant phase transitions make a SPLash. *Cell*, 138(4):625–627, 2009.

Fraser, Christopher M; Thompson, Michael G; Shirley, Amber M; Ralph, John; Schoenherr, Jessica A; Sinlapadech, Taksina; Hall, Mark C, and Chapple, Clint. Related arabidopsis serine carboxypeptidase-like sinapoylglucose acyltransferases display distinct but overlapping substrate specificities. *Plant Physiology*, 144(4):1986–1999, 2007.

Gallagher, Thomas L and Gasser, Charles S. Independence and interaction of regions of the INNER NO OUTER protein in growth control during ovule development. *Plant Physiology*, 147(1):306–315, 2008.

Gardner, Timothy S and Faith, Jeremiah J. Reverse-engineering transcription control networks. *Physics of Life Reviews*, 2(1):65–88, 2005.

Gardner, Timothy S; di Bernardo, Diego; Lorenz, David, and Collins, James J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science Signalling*, 301(5629):102, 2003.

Gat-Viks, Irit and Shamir, Ron. Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Research*, 17(3):358–367, 2007.

Gechev, Tsanko; Minkov, Ivan, and Hille, Jacques. Hydrogen peroxide-induced cell death in arabidopsis: Transcriptional and mutant analysis reveals a role of an oxoglutarate-dependent dioxygenase gene in the cell death process. *IUBMB Life*, 57(3):181–188, 2008.

Genoud, Thierry; Santa Cruz, Marcela B Trevino, and Métraux, Jean-Pierre. Numeric simulation of plant signaling networks. *Plant Physiology*, 126(4):1430–1437, 2001.

Glynn, Steven E; Baker, Patrick J; Sedelnikova, Svetlana E; Davies, Claire L; Eadsforth, Thomas C; Levy, Colin W; Rodgers, H Fiona; Blackburn, G Michael; Hawkes, Timothy R; Viner, Russell, and others, . Structure and mechanism of imidazoleglycerol-phosphate dehydratase. *Structure*, 13(12):1809–1817, 2005.

Gómez-Mena, Concepción; de Folter, Stefan; Costa, Maria Manuela R; Angenent, Gerco C, and Sablowski, Robert. Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. *Development*, 132(3):429–438, 2005.

Gonzalez, Antonio; Zhao, Mingzhe; Leavitt, John M, and Lloyd, Alan M. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in arabidopsis seedlings. *The Plant Journal*, 53(5):814–827, 2007.

Grandi, Valentina; Gregis, Veronica, and Kater, Martin M. Uncovering genetic and molecular interactions among floral meristem identity genes in arabidopsis thaliana. *The Plant Journal*, 2012.

Gregoretti, Francesco; Belcastro, Vincenzo; Di Bernardo, Diego, and Oliva, Gennaro. A parallel implementation of the network identification by multiple regression (nir) algorithm to reverse-engineer regulatory gene networks. *PloS One*, 5(4):e10179, 2010.

Groszmann, Michael; Paicu, Teodora; Alvarez, John P; Swain, Steve M, and Smyth, David R. SPATULA and ALCATRAZ, are partially redundant, functionally diverging bHLH genes required for arabidopsis gynoecium and fruit development. *The Plant Journal*, 68(5):816–829, 2011.

Gupta, Kapuganti J; Shah, Jay K; Brotman, Yariv; Jahnke, Kathrin; Willmitzer, Lothar; Kaiser, Werner M; Bauwe, Hermann, and Igamberdiev, Abir U. Inhibition of aconitase by nitric oxide leads to induction of the alternative oxidase and to a shift of metabolism towards biosynthesis of amino acids. *Journal of Experimental Botany*, 63(4):1773–1784, 2012.

Haibe-Kains, Benjamin; Olsen, Catharina; Djebbari, Amira; Bontempi, Gianluca; Correll, Mick; Bouton, Christopher, and Quackenbush, John. Predictive networks: a flexible, open source, web application for integration and analysis of human gene networks. *Nucleic Acids Research*, 40(D1):D866–D875, 2012.

Hamon, M; Bourgoin, S; Artaud, F; El Mestikawy, S, and others, . The respective roles of tryptophan uptake and tryptophan hydroxylase in the regulation of serotonin synthesis in the central nervous system. *Journal de Physiologie*, 77(2-3):269, 1981.

Han, Soon-Ki; Song, Ju-Dong; Noh, Yoo-Sun, and Noh, Bosl. Role of plant CBP/p300-like genes in the regulation of flowering time. *The Plant Journal*, 49(1):103–114, 2006.

Harborne, Jeffrey B and Williams, Christine A. Advances in flavonoid research since 1992. *Phytochemistry*, 55 (6):481–504, 2000.

Harding, Scott A; Leshkevich, Jacqueline; Chiang, Vincent L, and Tsai, Chung-Jui. Differential substrate inhibition couples kinetically distinct 4-coumarate: coenzyme a ligases with spatially distinct metabolic roles in quaking aspen. *Plant Physiology*, 128(2):428–438, 2002.

Hashimoto, Ronaldo F; Kim, Seungchan; Shmulevich, Ilya; Zhang, Wei; Bittner, Michael L, and Dougherty, Edward R. Growing genetic regulatory networks from seed genes. *Bioinformatics*, 20(8):1241–1247, 2004.

Hecker, Michael; Lambeck, Sandro; Toepfer, Susanne; van Someren, Eugene, and Guthke, Reinhard. Gene regulatory network inference: Data integration in dynamic modelsŮa. *Biosystems*, 96:86–103, 2009.

Heckerman, David; Geiger, Dan, and Chickering, David M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

Hemm, Matthew R; Rider, Stanley D; Ogas, Joseph; Murry, Daryl J, and Chapple, Clint. Light induces phenylpropanoid metabolism in arabidopsis roots. *The Plant Journal*, 38(5):765–778, 2004.

Herridge, Rowan Paul. *Molecular genetic approaches to understanding and manipulating seed development in Arabidopsis thaliana*. PhD thesis, University of Otago, 2012.

Hoang, My Hanh Thi; Nguyen, Xuan Canh; Lee, Kyunghee; Kwon, Young Sang; Pham, Huyen Trang Thi; Park, Hyeong Cheol; Yun, Dae-Jin; Lim, Chae Oh, and Chung, Woo Sik. Phosphorylation by atMPK6 is required for the biological function of atMYB41 in *Arabidopsis*. *Biochemical and Biophysical Research Communications*, 422(1):181–186, 2012.

Hodges, Andrew P; Dai, Dongjuan; Xiang, Zuoshuang; Woolf, Peter; Xi, Chuanwu, and He, Yongqun. Bayesian network expansion identifies new ros and biofilm regulators. *PLoS One*, 5(3):e9513, 2010.

Höfer, Rene; Briesen, Isabel; Beck, Martina; Pinot, Franck; Schreiber, Lukas, and Franke, Rochus. The arabidopsis cytochrome P450 CYP86A1 encodes a fatty acid ω-hydroxylase involved in suberin monomer biosynthesis. *Journal of Experimental Botany*, 59(9):2347–2360, 2008.

Hoffmann, Laurent; Besseau, Sébastien; Geoffroy, Pierrette; Ritzenthaler, Christophe; Meyer, Denise; Lapierre, Catherine; Pollet, Brigitte, and Legrand, Michel. Silencing of hydroxycinnamoyl-coenzyme a shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *The Plant Cell Online*, 16(6):1446–1465, 2004.

Holst, Kerstin; Schmülling, Thomas, and Werner, Tomáš. Enhanced cytokinin degradation in leaf primordia of transgenic *Arabidopsis* plants reduces leaf size and shoot organ primordia formation. *Journal of Plant Physiology*, 168(12):1328–1334, 2011.

Horák, Jakub; Grefen, Christopher; Berendzen, Kenneth W; Hahn, Achim; Stierhof, York-Dieter; Stadelhofer, Bettina; Stahl, Mark; Koncz, Csaba, and Harter, Klaus. The *Arabidopsis thaliana* response regulator ARR22 is a putative AHP phospho-histidine phosphatase expressed in the chalaza of developing seeds. *BMC Plant Biology*, 8(1):77, 2008.

Hornitschek, Patricia; Kohnen, Markus V; Lorrain, Séverine; Rougemont, Jacques; Ljung, Karin; López-Vidriero, Irene; Franco-Zorrilla, José M; Solano, Roberto; Trevisan, Martine; Pradervand, Sylvain, and others, . Phytochrome interacting factors 4 and 5 control seedling growth in changing light conditions by directly controlling auxin signaling. *The Plant Journal*, 2012.

Hsu, Hsing-Fun; Huang, Chih-Hsiang; Chou, Lu-Tung, and Yang, Chang-Hsien. Ectopic expression of an orchid (oncidium gower ramsey) AGL6-like gene promotes flowering by activating flowering time genes in *Arabidopsis thaliana*. *Plant and Cell Physiology*, 44(8):783–794, 2003.

Hu, Jinguo and Vick, Brady A. Target region amplification polymorphism: a novel marker technique for plant genotyping. *Plant Molecular Biology Reporter*, 21(3):289–294, 2003.

Hu, Wei; Wang, Yixing; Bowers, Christian, and Ma, Hong. Isolation, sequence analysis, and expression studies of florally expressed cdnas in arabidopsis. *Plant Molecular Biology*, 53(4):545–563, 2003.

Huang, Jun; Bhinu, V-S; Li, Xiang; Dallal Bashi, Zafer; Zhou, Rong, and Hannoufa, Abdelali. Pleiotropic changes in arabidopsis f5h and sct mutants revealed by large-scale gene expression and metabolite analysis. *Planta*, 230 (5):1057–1069, 2009.

Huang, Yihua; Baxter, Richard; Smith, Barbara S; Partch, Carrie L; Colbert, Christopher L, and Deisenhofer, Johann. Crystal structure of cryptochrome 3 from *Arabidopsis thaliana* and its implications for photolyase activity. *Proceedings of the National Academy of Sciences*, 103(47):17701–17706, 2006.

Hunt, Emma; Gattolin, Stefano; Newbury, H John; Bale, Jeffrey S; Tseng, Hua-Ming; Barrett, David A, and Pritchard, Jeremy. A mutation in amino acid permease aap6 reduces the amino acid content of the arabidopsis sieve elements but leaves aphid herbivores unaffected. *Journal of Experimental Botany*, 61(1):55–64, 2010.

Huq, Enamul and Quail, Peter H. PIF4, a phytochrome-interacting bHLH factor, functions as a negative regulator of phytochrome B signaling in arabidopsis. *The EMBO Journal*, 21(10):2441–2450, 2002.

Ihmels, Jan; Friedlander, Gilgi; Bergmann, Sven; Sarig, Ofer; Ziv, Yaniv; Barkai, Naama, and others, . Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31(4):370–378, 2002.

Irish, Vivian F. The flowering of arabidopsis flower development. *The Plant Journal*, 61(6):1014–1028, 2010.

Ishida, Tetsuya; Hattori, Sayoko; Sano, Ryosuke; Inoue, Kayoko; Shirano, Yumiko; Hayashi, Hiroaki; Shibata, Daisuke; Sato, Shusei; Kato, Tomohiko; Tabata, Satoshi, and others, . Arabidopsis TRANSPARENT TESTA GLABRA2 is directly regulated by R2R3 MYB transcription factors and is involved in regulation of GLABRA2 transcription in epidermal differentiation. *The Plant Cell Online*, 19(8):2531–2543, 2007.

Iturriaga, Gabriel; Suárez, Ramón, and Nova-Franco, Barbara. Trehalose metabolism: from osmoprotection to signaling. *International Journal of Molecular Sciences*, 10(9):3793–3810, 2009.

Jack, Thomas. Relearning our ABCs: new twists on an old model. *Trends in Plant Science*, 6(7):310–316, 2001.

Jansen, Mickel LA; Diderich, Jasper A; Mashego, Mlawule; Hassane, Adham; de Winde, Johannes H; Daran-Lapujade, Pascale, and Pronk, Jack T. Prolonged selection in aerobic, glucose-limited chemostat cultures of *Saccharomyces cerevisiae* causes a partial loss of glycolytic capacity. *Microbiology*, 151(5):1657–1669, 2005.

Jiang, Ling; Yuan, Li; Xia, Ming, and Makaroff, Christopher A. Proper levels of the arabidopsis cohesion establishment factor CTF7 are essential for embryo and megagametophyte, but not endosperm, development. *Plant Physiology*, 154(2):820–832, 2010.

Jun, JiHyung; Fiume, Elisa; Roeder, Adrienne HK; Meng, Ling; Sharma, Vijay K; Osmont, Karen S; Baker, Catherine; Ha, Chan Man; Meyerowitz, Elliot M; Feldman, Lewis J, and others, . Comprehensive analysis of CLE polypeptide signaling gene expression and overexpression activity in arabidopsis. *Plant Physiology*, 154 (4):1721–1736, 2010.

Kalisch, Markus and Bühlmann, Peter. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

Kalisch, Markus; Mächler, Martin; Colombo, Diego; Maathuis, Marloes H, and Bühlmann, Peter. Causal inference using graphical models with the r package pcalg. *Preprint, available at http://cran. rproject. org/web/packages/pcalg/vignettes/pcalgDoc. pdf*, 2010.

Kang, Jungen; Zhang, Guoyu; Bonnema, Guusje; Fang, Zhiyuan, and Wang, Xiaowu. Global analysis of gene expression in flower buds of ms-cd1 *Brassica oleracea* conferring male sterility by using an arabidopsis microarray. *Plant Molecular Biology*, 66(1):177–192, 2008.

Kauffman, Stuart A. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, 1969.

Kaufmann, Kerstin; Muino, Jose M; Jauregui, Ruy; Airoldi, Chiara A; Smaczniak, Cezary; Krajewski, Pawel, and Angenent, Gerco C. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the arabidopsis flower. *PLoS Biology*, 7(4):e1000090, 2009.

Kenny, David A. Correlation and causality. *New York: Wiley, 1979*, 1, 1979.

Kim, Sung-Jin; Kim, Kye-Won; Cho, Man-Ho; Franceschi, Vincent R; Davin, Laurence B; Lewis, Norman G, and others, . Expression of cinnamyl alcohol dehydrogenases and their putative homologues during arabidopsis thaliana growth and development: Lessons for database annotations? *Phytochemistry*, 68(14):1957, 2007.

Kinoshita, Natsuko; Berr, Alexandre; Belin, Christophe; Chappuis, Richard; Nishizawa, Naoko K, and Lopez-Molina, Luis. Identification of growth insensitive to ABA3 (gia3), a recessive mutation affecting ABA signaling for the control of early post-germination growth in *Arabidopsis thaliana*. *Plant and Cell Physiology*, 51(2):239–251, 2010.

Klopffleisch, Karsten; Phan, Nguyen; Augustin, Kelsey; Bayne, Robert S; Booker, Katherine S; Botella, Jose R; Carpita, Nicholas C; Carr, Tyrell; Chen, Jin-Gui; Cooke, Thomas Ryan, and others, . Arabidopsis g-protein interactome reveals connections to cell wall carbohydrates and morphogenesis. *Molecular Systems Biology*, 7 (1), 2011.

Kohn, Kurt W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10(8):2703–2734, 1999.

Kohn, Kurt W; Aladjem, Mirit I; Weinstein, John N, and Pommier, Yves. Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. *Molecular Biology of the Cell*, 17(1):1–13, 2006.

Koornneef, Maarten; Alonso-Blanco, Carlos; Peeters, Anton JM, and Soppe, Wim. Genetic control of flowering time in arabidopsis. *Annual Review of Plant Biology*, 49(1):345–370, 1998.

Kosarev, Peter; Mayer, KF; Hardtke, Christian S, and others, . Evaluation and classification of ring-finger domains encoded by the arabidopsis genome. *Genome Biology*, 3(4):0016–1, 2002.

Kotera, Masaaki; Yamanishi, Yoshihiro; Moriya, Yuki; Kanehisa, Minoru, and Goto, Susumu. GENIES: gene network inference engine based on supervised analysis. *Nucleic Acids Research*, 40(W1):W162–W167, 2012.

Krizek, Beth A; Prost, Valerie, and Macias, Anthony. AINTEGUMENTA promotes petal identity and acts as a negative regulator of AGAMOUS. *The Plant Cell Online*, 12(8):1357–1366, 2000.

Larsen, Paul B; Geisler, Matt JB; Jones, Carol A; Williams, Kelly M, and Cancel, Jesse D. ALS3 encodes a phloem-localized abc transporter-like protein that is required for aluminum tolerance in arabidopsis. *The Plant Journal*, 41(3):353–363, 2004.

Lauria, Mario and di Bernardo, Diego. Reconstructing gene networks using gene expression profiles. *Cancer Systems Biology*, 32:35, 2010.

Lechner, E; Leonhardt, N; Eisler, H; Parmentier, Y; Alioua, M; Jacquet, H; Leung, J, and Genschik, P. MATH-/BTB CRL3 receptors target the homeodomain-leucine zipper ATHB6 to modulate abscisic acid signaling. *Developmental Cell*, 21(6):1116–1128, 2011.

Lee, Homin K; Hsu, Amy K; Sajdak, Jon; Qin, Jie, and Pavlidis, Paul. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14(6):1085–1094, 2004.

Lee, Ji-Young; Baum, Stuart F; Oh, Sang-Hun; Jiang, Cai-Zhong; Chen, Jen-Chih, and Bowman, John L. Recruitment of CRABS CLAW to promote nectary development within the eudicot clade. *Development*, 132 (22):5021–5032, 2005.

Lee, Keun Pyo; Piskurewicz, Urszula; Turečková, Veronika; Strnad, Miroslav, and Lopez-Molina, Luis. A seed coat bedding assay shows that RGL2-dependent release of abscisic acid by the endosperm controls embryo growth in arabidopsis dormant seeds. *Proceedings of the National Academy of Sciences*, 107(44):19108–19113, 2010.

Lenhard, Michael and Laux, Thomas. Stem cell homeostasis in the arabidopsis shoot meristem is regulated by intercellular movement of CLAVATA3 and its sequestration by CLAVATA1. *Development*, 130(14):3163–3173, 2003.

Lewin, Benjamin and Dover, Gabby. *Genes v*, volume 299. Oxford University Press New York, 1994.

Li, Pinghua; Ma, Shisong, and Bohnert, Hans J. Coexpression characteristics of trehalose-6-phosphate phosphatase subfamily genes reveal different functions in a network context. *Physiologia Plantarum*, 133(3):544–556, 2008.

Li, Song Feng; Milliken, Olga Nicolaou; Pham, Hanh; Seyit, Reg; Napoli, Ross; Preston, Jeremy; Koltunow, Anna M, and Parish, Roger W. The arabidopsis MYB5 transcription factor regulates mucilage synthesis, seed coat development, and trichome morphogenesis. *The Plant Cell Online*, 21(1):72–89, 2009.

Li, Tao; Hu, Yajun; Hao, Zhipeng; Li, Hong, and Chen, Baodong. Aquaporin genes gintAQPF1 and gintAQPF2 from glomus intraradices contribute to plant drought tolerance. *Plant Signaling & Behavior*, 8(5):e24030, 2013.

Li, Xiao-Chuan; Zhu, Jun; Yang, Jun; Zhang, Guo-Rui; Xing, Wei-Feng; Zhang, Sen, and Yang, Zhong-Nan. Glycerol-3-phosphate acyltransferase 6 (GPAT6) is important for tapetum development in arabidopsis and plays multiple roles in plant fertility. *Molecular Plant*, 5(1):131–142, 2012.

Li, Xing Guo; Su, Ying Hua; Zhao, Xiang Yu; Li, Wei; Gao, Xin Qi, and Zhang, Xian Sheng. Cytokinin overproduction-caused alteration of flower development is partially mediated by CUC2 and CUC3 in arabidopsis. *Gene*, 450(1):109–120, 2010.

Liang, Shoudan; Fuhrman, Stefanie; Somogyi, Roland, and others, . REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific symposium on biocomputing*, 3(18-29):2, 1998.

Lin, Rongcheng and Wang, Haiyang. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiology*, 136(4):4010–4022, 2004.

Liu, Li; Qi, Hongying; Wang, Jianquan, and Lin, Haifan. PAPI, a novel TUDOR-domain protein, complexes with AGO3, ME31B and TRAL in the nuage to silence transposition. *Development*, 138(9):1863–1873, 2011.

Lockhart, David J; Dong, Helin; Byrne, Michael C; Follettie, Maximillian T; Gallo, Michael V; Chee, Mark S; Mittmann, Michael; Wang, Chunwei; Kobayashi, Michiko; Norton, Heidi, and others, . Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.

Logemann, Elke; Tavernaro, Annette; Schulz, Wolfgang; Somssich, Imre E, and Hahlbrock, Klaus. UV light selectively coinduces supply pathways from primary metabolism and flavonoid secondary product formation in parsley. *Proceedings of the National Academy of Sciences*, 97(4):1903–1907, 2000.

Ma, Xuan; Feng, Baomin, and Ma, Hong. AMS-dependent and independent regulation of anther transcriptome and comparison with those affected by other *Arabidopsis* anther genes. *BMC Plant Biology*, 12(1):23, 2012.

Maizel, Alexis; Busch, Maximilian A; Tanahashi, Takako; Perkovic, Josip; Kato, Masahiro; Hasebe, Mitsuyasu, and Weigel, Detlef. The floral regulator LEAFY evolves by substitutions in the dna binding domain. *Science Signalling*, 308(5719):260, 2005.

Mangeon, Amanda; Magioli, Claudia; Tarré, Érika; Cardeal, Vanessa; Araujo, Cristina; Falkenbach, Erica; Rocha, Carla Andréa Benício; Rangel-Lima, Camila, and Sachetto-Martins, Gilberto. The tissue expression pattern of the atgrp5 regulatory region is controlled by a combination of positive and negative elements. *Plant Cell Reports*, 29(5):461–471, 2010.

Marbach, Daniel; Schaffter, Thomas; Mattiussi, Claudio, and Floreano, Dario. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16 (2):229–239, 2009.

Margolin, Adam A; Nemenman, Ilya; Basso, Katia; Wiggins, Chris; Stolovitzky, Gustavo; Favera, Riccardo D, and Califano, Andrea. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

Margulies, Marcel; Egholm, Michael; Altman, William E; Attiya, Said; Bader, Joel S; Bemben, Lisa A; Berka, Jan; Braverman, Michael S; Chen, Yi-Ju; Chen, Zhoutao, and others, . Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

Marín-Rodríguez, M Celia; Orchard, John, and Seymour, Graham B. Pectate lyases, cell wall degradation and fruit softening. *Journal of Experimental Botany*, 53(377):2115–2119, 2002.

Marinova, Krasimira; Pourcel, Lucille; Weder, Barbara; Schwarz, Michael; Barron, Denis; Routaboul, Jean-Marc; Debeaujon, Isabelle, and Klein, Markus. The arabidopsis MATE transporter TT12 acts as a vacuolar flavonoid/h+-antiporter active in proanthocyanidin-accumulating cells of the seed coat. *The Plant Cell Online*, 19(6):2023–2038, 2007.

Marioni, John C; Mason, Christopher E; Mane, Shrikant M; Stephens, Matthew, and Gilad, Yoav. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9): 1509–1517, 2008.

Matsumoto, Noritaka and Okada, Kiyotaka. A homeobox gene, PRESSED FLOWER, regulates lateral axis-dependent development of arabidopsis flowers. *Genes & Development*, 15(24):3355–3364, 2001.

Medintz, Igor L; Vora, Gary J; Rahbar, Amir M, and Thach, Dzung C. Transcript and proteomic analyses of wild-type and *gpa2* mutant *Saccharomyces cerevisiae* strains suggest a role for glycolytic carbon source sensing in pseudohyphal differentiation. *Molecular BioSystem*, 3(9):623–634, 2007.

Meyer, Patrick E; Kontos, Kevin; Lafitte, Frederic, and Bontempi, Gianluca. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007, 2007.

Mizutani, Masaharu and Ohta, Daisaku. Diversification of P450 genes during land plant evolution. *Annual Review of Plant Biology*, 61:291–315, 2010.

Molas, Maria Lia; Kiss, John Z, and Correll, Melanie J. Gene profiling of the red light signalling pathways in roots. *Journal of Experimental Botany*, 57(12):3217–3229, 2006.

Mølhøj, Michael; Ulvskov, Peter, and Dal Degan, Florence. Characterization of a functional soluble form of a *Brassica napus* membrane-anchored endo-1, 4-β-glucanase heterologously expressed in *Pichia pastoris*. *Plant Physiology*, 127(2):674–684, 2001.

Morris, Marilyn E and Zhang, Shuzhong. Flavonoid–drug interactions: effects of flavonoids on ABC transporters. *Life Sciences*, 78(18):2116–2130, 2006.

Mostafavi, Sara; Ray, Debajyoti; Warde-Farley, David; Grouios, Chris; Morris, Quaid, and others, . GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(Suppl 1):S4, 2008.

Mougiou, Niki; Poulios, Stylianos; Kaldis, Athanasios, and Vlachonasios, Konstantinos E. *Arabidopsis thaliana* TBP-associated factor 5 is essential for plant growth and development. *Molecular Breeding*, 30(1):355–366, 2012.

Müller, Joachim; Aeschbacher, Roger A; Wingler, Astrid; Boller, Thomas, and Wiemken, Andres. Trehalose and trehalase in arabidopsis. *Plant Physiology*, 125(2):1086–1093, 2001.

Murmu, Jhadeswar; Bush, Michael J; DeLong, Catherine; Li, Shutian; Xu, Mingli; Khan, Madiha; Malcolmson, Caroline; Fobert, Pierre R; Zachgo, Sabine, and Hepworth, Shelley R. Arabidopsis basic leucine-zipper transcription factors TGA9 and TGA10 interact with floral glutaredoxins ROXY1 and ROXY2 and are redundantly required for anther development. *Plant Physiology*, 154(3):1492–1504, 2010.

Myers, Chad L; Robson, Drew; Wible, Adam; Hibbs, Matthew A; Chiriac, Camelia; Theesfeld, Chandra L; Dolinski, Kara; Troyanskaya, Olga G, and others, . Discovery of biological networks from diverse functional genomic data. *Genome Biology*, 6(13):R114, 2005.

Nakaune, Satoru; Yamada, Kenji; Kondo, Maki; Kato, Tomohiko; Tabata, Satoshi; Nishimura, Mikio, and Hara-Nishimura, Ikuko. A vacuolar processing enzyme, δVPE, is involved in seed coat formation at the early stage of seed development. *The Plant Cell Online*, 17(3):876–887, 2005.

Nemenman, Ilya; Hlavacek, William S; Unkefer, Pat J; Unkefer, Clifford J; Wall, Michael E, and others, . Reconstruction of metabolic networks from high-throughput metabolite profiling data. *Annals of the New York Academy of Sciences*, 1115(1):102–115, 2007.

Nishikawa, Masahiro; Hosokawa, Kenta; Ishiguro, Mai; Minamioka, Hiroki; Tamura, Kentaro; Hara-Nishimura, Ikuko; Takahashi, Yohei; Shimazaki, Ken-ichiro, and Imai, Hiroyuki. Degradation of sphingoid long-chain base 1-phosphates (LCB-1Ps): functional characterization and expression of atDPL1 encoding LCB-1P lyase involved in the dehydration stress response in arabidopsis. *Plant and Cell Physiology*, 49(11):1758–1763, 2008.

Nonomura, Ken-Ichi; Miyoshi, Kazumaru; Eiguchi, Mitsugu; Suzuki, Tadzunu; Miyao, Akio; Hirochika, Hirohiko, and Kurata, Nori. The MSP1 gene is necessary to restrict the number of cells entering into male and female sporogenesis and to initiate anther wall formation in rice. *The Plant Cell Online*, 15(8):1728–1739, 2003.

Nutricati, Eliana; Miceli, Antonio; Blando, Federica, and De Bellis, Luigi. Characterization of two *Arabidopsis thaliana* glutathione s-transferases. *Plant Cell Reports*, 25(9):997–1005, 2006.

Olmedo-Monfil, Vianey; Durán-Figueroa, Noé; Arteaga-Vázquez, Mario; Demesa-Arévalo, Edgar; Autran, Daphné; Grimanelli, Daniel; Slotkin, R Keith; Martienssen, Robert A, and Vielle-Calzada, Jean-Philippe. Control of female gamete formation by a small RNA pathway in arabidopsis. *Nature*, 464(7288):628–632, 2010.

Olsen, Kristine M; Lea, Unni S; Slimestad, Rune; Verheul, Michel, and Lillo, Cathrine. Differential expression of four *Arabidopsis* PAL genes; PAL1 and PAL2 have functional specialization in abiotic environmental-triggered flavonoid synthesis. *Journal of Plant Physiology*, 165(14):1491–1499, 2008.

Onda, Yayoi; Yagi, Yusuke; Saito, Yukiko; Takenaka, Nobuhiro, and Toyoshima, Yoshinori. Light induction of arabidopsis SIG1 and SIG5 transcripts in mature leaves: differential roles of cryptochrome 1 and cryptochrome 2 and dual function of sig5 in the recognition of plastid promoters. *The Plant Journal*, 55(6):968–978, 2008.

Pearl, Judea. Causality: models, reasoning, and inference. *IIE Transactions*, 34(6):583–589, 2002.

Pe'er, Dana. Bayesian network analysis of signaling networks: a primer. *Science Signalling*, 2005(281):pl4, 2005.

Peer, Wendy Ann and Murphy, Angus S. Flavonoids and auxin transport: modulators or regulators? *Trends in Plant Science*, 12(12):556–563, 2007.

Penfold, Christopher A and Wild, David L. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

Posé, David; Castanedo, Itziar; Borsani, Omar; Nieto, Benjamín; Rosado, Abel; Taconnat, Ludivine; Ferrer, Albert; Dolan, Liam; Valpuesta, Victoriano, and Botella, Miguel A. Identification of the arabidopsis dry2/sqe1-5 mutant reveals a central role for sterols in drought tolerance and regulation of reactive oxygen species. *The Plant Journal*, 59(1):63–76, 2009.

Posé, David; Yant, Levi, and Schmid, Markus. The end of innocence: flowering networks explode in complexity. *Current Opinion in Plant Biology*, 2011.

Preston, Jeremy; Wheeler, Janet; Heazlewood, Joshua; Li, Song Feng, and Parish, Roger W. AtMYB32 is required for normal pollen development in *Arabidopsis thaliana*. *The Plant Journal*, 40(6):979–995, 2004.

Qi, Tiancong; Song, Susheng; Ren, Qingcuo; Wu, Dewei; Huang, Huang; Chen, Yan; Fan, Meng; Peng, Wen; Ren, Chunmei, and Xie, Daoxin. The jasmonate-ZIM-domain proteins interact with the WD-repeat/bHLH/MYB complexes to regulate jasmonate-mediated anthocyanin accumulation and trichome initiation in *Arabidopsis thaliana*. *The Plant Cell Online*, 23(5):1795–1814, 2011a.

Qi, XiaoLi; Jiang, Yao; Tang, Fang; Wang, MinJie; Hu, JianJun; Zhao, ShuTang; Sha, Wei, and Lu, MengZhu. An *Arabidopsis thaliana* (ler) inbred line afdl exhibiting abnormal flower development mainly caused by reduced AP1 expression. *Chinese Science Bulletin*, 56(1):39–47, 2011b.

Quesada, VICTOR; Dean, CAROLINE; Simpson, GORDON G, and others, . Regulated RNA processing in the control of arabidopsis flowering. *International Journal of Developmental Biology*, 49(5/6):773, 2005.

Ramsey, Joseph; Zhang, Jiji, and Spirtes, Peter L. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

Ranocha, Philippe; Denancé, Nicolas; Vanholme, Ruben; Freydier, Amandine; Martinez, Yves; Hoffmann, Laurent; Köhler, Lothar; Pouzet, Cécile; Renou, Jean-Pierre; Sundberg, Björn, and others, . Walls are thin 1 (WAT1), an arabidopsis homolog of medicago truncatula NODULIN21, is a tonoplast-localized protein required for secondary wall formation in fibers. *The Plant Journal*, 63(3):469–483, 2010.

Riemann, M; Gutjahr, C; Korte, A; Danger, B; Muramatsu, T; Bayer, U; Waller, F; Furuya, M, and Nick, P. GER1, a GDSL motif-encoding gene from rice is a novel early light-and jasmonate-induced gene. *Plant Biology*, 9(1):32–40, 2008.

Rieu, Ivo; Ruiz-Rivero, Omar; Fernandez-Garcia, Nieves; Griffiths, Jayne; Powers, Stephen J; Gong, Fan; Linhartova, Terezie; Eriksson, Sven; Nilsson, Ove; Thomas, Stephen G, and others, . The gibberellin biosynthetic genes atGA20ox1 and atGA20ox2 act, partially redundantly, to promote growth and development throughout the arabidopsis life cycle. *The Plant Journal*, 53(3):488–504, 2007.

Rijpkema, Anneke S; Zethof, Jan; Gerats, Tom, and Vandenbussche, Michiel. The petunia AGL6 gene has a SEPALLATA-like function in floral patterning. *The Plant Journal*, 60(1):1–9, 2009.

Ro, Dae-Kyun; Ehlting, Jürgen; Keeling, Christopher I; Lin, Roy; Mattheus, Nathalie, and Bohlmann, Jörg. Microarray expression profiling and functional characterization of *AtTPS* genes: Duplicated *Arabidopsis thaliana* sesquiterpene synthase genes *At4g13280* and *At4g13300* encode root-specific and wound-inducible (*Z*)-γ-bisabolene synthases. *Archives of Biochemistry and Biophysics*, 448(1):104–116, 2006.

Rohde, Antje; Morreel, Kris; Ralph, John; Goeminne, Geert; Hostyn, Vanessa; De Rycke, Riet; Kushnir, Sergej; Van Doorsselaere, Jan; Joseleau, Jean-Paul; Vuylsteke, Marnik, and others, . Molecular phenotyping of the pal1 and pal2 mutants of *Arabidopsis thaliana* reveals far-reaching consequences on phenylpropanoid, amino acid, and carbohydrate metabolism. *The Plant Cell Online*, 16(10):2749–2771, 2004.

Romanel, Elisson; Das, Pradeep; Amasino, Richard M; Traas, Jan; Meyerowitz, Elliot, and Alves-Ferreira, Marcio. Reproductive meristem22 is a unique marker for the early stages of stamen development. *International Journal of Developmental Biology*, 55(6):657, 2011.

Romanel, Elisson AC; Schrago, Carlos G; Couñago, Rafael M; Russo, Claudia AM, and Alves-Ferreira, Márcio. Evolution of the B3 DNA binding superfamily: new insights into REM family gene diversification. *PloS One*, 4(6):e5791, 2009.

Routaboul, Jean-Marc; Dubos, Christian; Beck, Gilles; Marquis, Catherine; Bidzinski, Przemyslaw; Loudet, Olivier, and Lepiniec, Loïc. Metabolite profiling and quantitative genetics of natural variation for flavonoids in arabidopsis. *Journal of Experimental Botany*, 63(10):3749–3764, 2012.

Ruegger, Max and Chapple, Clint. Mutations that reduce sinapoylmalate accumulation in arabidopsis thaliana define loci with diverse roles in phenylpropanoid metabolism. *Genetics*, 159(4):1741–1749, 2001.

Ryan, Ken G; Swinny, Ewald E; Markham, Kenneth R, and Winefield, Chris. Flavonoid gene expression and UV photoprotection in transgenic and mutant *Petunia* leaves. *Phytochemistry*, 59(1):23–32, 2002.

Sanchez-Corrales, Yara-Elena; Alvarez-Buylla, Elena R, and Mendoza, Luis. The *Arabidopsis thaliana* flower organ specification gene regulatory network determines a robust differentiation process. *Journal of Theoretical Biology*, 264(3):971–983, 2010.

Scheible, Wolf-Rüdiger; Morcuende, Rosa; Czechowski, Tomasz; Fritz, Christina; Osuna, Daniel; Palacios-Rojas, Natalia; Schindelasch, Dana; Thimm, Oliver; Udvardi, Michael K, and Stitt, Mark. Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of arabidopsis in response to nitrogen. *Plant Physiology*, 136(1):2483–2499, 2004.

Schena, Mark; Shalon, Dari; Heller, Renu; Chai, Andrew; Brown, Patrick O, and Davis, Ronald W. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences*, 93(20):10614–10619, 1996.

Schmid, Markus; Uhlenhaut, N Henriette; Godard, François; Demar, Monika; Bressan, Ray; Weigel, Detlef, and Lohmann, Jan U. Dissection of floral induction pathways using global expression analysis. *Development*, 130 (24):6001–6012, 2003.

Schwarz, Gideon. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Semchuk, Nadia M; Lushchak, Oleh V; Falk, Jon; Krupinska, Karin, and Lushchak, Volodymyr I. Inactivation of genes, encoding tocopherol biosynthetic pathway enzymes, results in oxidative stress in outdoor grown *Arabidopsis thaliana*. *Plant Physiology and Biochemistry*, 47(5):384–390, 2009.

Seo, Eunjoo; Lee, Horim; Jeon, Jin; Park, Hanna; Kim, Jungmook; Noh, Yoo-Sun, and Lee, Ilha. Crosstalk between cold response and flowering in arabidopsis is mediated through the flowering-time gene SOC1 and its upstream negative regulator FLC. *The Plant Cell Online*, 21(10):3185–3197, 2009.

Seo, Pil Joon; Hong, Shin-Young; Kim, Sang-Gyu, and Park, Chung-Mo. Competitive inhibition of transcription factors by small interfering peptides. *Trends in Plant Science*, 16(10):541–549, 2011.

Serra, R and Villani, M. Modelling bacterial degradation of organic compounds with genetic networks. *Journal of Theoretical Biology*, 189(1):107–119, 1997.

Serrano, Mario and Guzmán, Plinio. Isolation and gene expression analysis of *Arabidopsis thaliana* mutants with constitutive expression of ATL2, an early elicitor-response RING-H2 zinc-finger gene. *Genetics*, 167(2): 919–929, 2004.

Shi, Chun-Lin; Stenvik, Grethe-Elisabeth; Vie, Ane Kjersti; Bones, Atle M; Pautot, Véronique; Proveniers, Marcel; Aalen, Reidunn B, and Butenko, Melinka A. Arabidopsis class i KNOTTED-like homeobox proteins act downstream in the iDA-HAE/HSL2 floral abscission signaling pathway. *The Plant Cell Online*, 23(7): 2553–2567, 2011a.

Shi, Jian Xin; Malitsky, Sergey; De Oliveira, Sheron; Branigan, Caroline; Franke, Rochus B; Schreiber, Lukas, and Aharoni, Asaph. SHINE transcription factors act redundantly to pattern the archetypal surface of arabidopsis flower organs. *PLoS Genetics*, 7(5):e1001388, 2011b.

Shymko, RM; De Meyts, P, and Thomas, R. Logical analysis of timing-dependent receptor signalling specificity: application to the insulin receptor metabolic and mitogenic signalling pathways. *Biochemical Journal*, 326(Pt 2):463, 1997.

Siddique, Shahid; Wieczorek, Krzysztof; Szakasits, Dagmar; Kreil, David P, and Bohlmann, Holger. The promoter of a plant defensin gene directs specific expression in nematode-induced syncytia in arabidopsis roots. *Plant Physiology and Biochemistry*, 49(10):1100–1107, 2011.

Siefers, Nicholas; Dang, Kristen K; Kumimoto, Roderick W; Bynum IV, William Edwards; Tayrose, Gregory, and Holt III, Ben F. Tissue-specific expression patterns of arabidopsis NF-Y transcription factors suggest potential for extensive combinatorial complexity. *Plant Physiology*, 149(2):625–641, 2009.

Singh, Jatinder; Kumar, Deept; Ramakrishnan, Naren; Singhal, Vibha; Jervis, Jody; Garst, James F; Slaughter, Stephen M; DeSantis, Andrea M; Potts, Malcolm, and Helm, Richard F. Transcriptional response of *Saccharomyces cerevisiae* to desiccation and rehydration. *Applied and Environmental Microbiology*, 71(12):8752–8763, 2005.

Sliwinski, MK; White, MA; Maizel, Alexis; Weigel, Detlef, and Baum, DA. Evolutionary divergence of LFY function in the mustards *Arabidopsis thaliana* and *Leavenworthia crassa*. *Plant Molecular Biology*, 62(1): 279–289, 2006.

Smaczniak, Cezary; Immink, Richard GH; Muiño, Jose M; Blanvillain, Robert; Busscher, Marco; Busscher-Lange, Jacqueline; Dinh, QD Peter; Liu, Shujing; Westphal, Adrie H; Boeren, Sjef, and others, . Characterization of MADS-domain transcription factor complexes in arabidopsis flower development. *Proceedings of the National Academy of Sciences*, 109(5):1560–1565, 2012.

Smith, Erin N and Kruglyak, Leonid. Gene–environment interaction in yeast gene expression. *PLoS Biology*, 6 (4):e83, 2008.

Spirtes, Peter and Glymour, Clark. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

Spirtes, Peter; Glymour, Clark, and Scheines, Richard. *Causation, prediction, and search*, volume 81. MIT press, 2001.

Ståldal, Veronika; Cierlik, Izabela; Chen, Song; Landberg, Katarina; Baylis, Tammy; Myrenås, Mattias; Sundström, Jens F; Eklund, D Magnus; Ljung, Karin, and Sundberg, Eva. The *Arabidopsis thaliana* transcriptional activator STYLISH1 regulates genes affecting stamen development, cell expansion and timing of flowering. *Plant Molecular Biology*, pages 1–15, 2012.

Stark, Chris; Breitkreutz, Bobby-Joe; Reguly, Teresa; Boucher, Lorrie; Breitkreutz, Ashton, and Tyers, Mike. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539, 2006.

Steuer, Ralf; Kurths, Jürgen; Daub, Carstens O; Weise, Janko, and Selbig, Joachim. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2):S231–S240, 2002.

Stolovitzky, Gustavo; Monroe, DON, and Califano, Andrea. Dialogue on reverse-engineering assessment and methods. *Annals of the New York Academy of Sciences*, 1115(1):1–22, 2007.

Stolovitzky, Gustavo; Prill, Robert J, and Califano, Andrea. Lessons from the DREAM2 challenges. *Annals of the New York Academy of Sciences*, 1158(1):159–195, 2009.

Stracke, Ralf; Ishihara, Hirofumi; Huep, Gunnar; Barsch, Aiko; Mehrtens, Frank; Niehaus, Karsten, and Weisshaar, Bernd. Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. *The Plant Journal*, 50(4):660–677, 2007.

Swaminathan, Kankshita; Peterson, Kevin, and Jack, Thomas. The plant B3 superfamily. *Trends in Plant Science*, 13(12):647–655, 2008.

Swatek, Kirby N; Graham, Katherine; Agrawal, Ganesh K, and Thelen, Jay J. The 14-3-3 isoforms chi and epsilon differentially bind client proteins from developing arabidopsis seed. *Journal of Proteome Research*, 10 (9):4076–4087, 2011.

Takahashi, Naoki; Kuroda, Hirofumi; Kuromori, Takashi; Hirayama, Takashi; Seki, Motoaki; Shinozaki, Kazuo; Shimada, Hiroaki, and Matsui, Minami. Expression and interaction analysis of arabidopsis skp1-related genes. *Plant and Cell Physiology*, 45(1):83–91, 2004.

Takase, Tomoyuki; Nakazawa, Miki; Ishikawa, Akie; Manabe, Katsushi, and Matsui, Minami. DFL2, a new member of the arabidopsis GH3 gene family, is involved in red light-specific hypocotyl elongation. *Plant and Cell Physiology*, 44(10):1071–1080, 2003.

Takei, Kentaro; Ueda, Nanae; Aoki, Koh; Kuromori, Takashi; Hirayama, Takashi; Shinozaki, Kazuo; Yamaya, Tomoyuki, and Sakakibara, Hitoshi. AtIPT3 is a key determinant of nitrate-dependent cytokinin biosynthesis in arabidopsis. *Plant and Cell Physiology*, 45(8):1053–1062, 2004.

Tanay, Amos; Shamir, Ron, and others, . Computational expansion of genetic networks. *Bioinformatics-Oxford-*, 17:270–278, 2001.

Tantikanjana, Titima; Rizvi, Noreen; Nasrallah, Mikhail E, and Nasrallah, June B. A dual role for the s-locus receptor kinase in self-incompatibility and pistil development revealed by an *Arabidopsis rdr6* mutation. *The Plant Cell Online*, 21(9):2642–2654, 2009.

Tarantino, Delia; Petit, Jean-Michel; Lobreaux, Stephane; Briat, Jean-François; Soave, Carlo, and Murgia, Irene. Differential involvement of the IDRS cis-element in the developmental and environmental regulation of the atfer1 ferritin gene from arabidopsis. *Planta*, 217(5):709–716, 2003.

Tarutani, Yoshiaki; Morimoto, Takashi; Sasaki, Akiko; Yasuda, Michiko; Nakashita, Hideo; Yoshida, Shigeo; Yamaguchi, Isomaro, and Suzuki, Yoshihito. Molecular characterization of two highly homologous receptor-like kinase genes, RLK902 and RKL1, in *Arabidopsis thaliana*. *Bioscience, Biotechnology, and Biochemistry*, 68(9):1935–1941, 2004.

Tegner, Jesper; Yeung, MK Stephen; Hasty, Jeff, and Collins, James J. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Sciences*, 100(10):5944–5949, 2003.

Thévenin, Johanne; Pollet, Brigitte; Letarnec, Bruno; Saulnier, Luc; Gissot, Lionel; Maia-Grondard, Alessandra; Lapierre, Catherine, and Jouanin, Lise. The simultaneous repression of CCR and CAD, two enzymes of the lignin biosynthetic pathway, results in sterility and dwarfism in *Arabidopsis thaliana*. *Molecular Plant*, 4(1): 70–82, 2011.

Tholl, Dorothea; Chen, Feng; Petri, Jana; Gershenzon, Jonathan, and Pichersky, Eran. Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from arabidopsis flowers. *The Plant Journal*, 42(5):757–771, 2005.

Thompson, Elinor P; Davies, Julia M, and Glover, Beverley J. Identifying the transporters of different flavonoids in plants. *Plant Signaling & Behavior*, 5(7):860–863, 2010.

Thors, Lina; Belghiti, M, and Fowler, CJ. Inhibition of fatty acid amide hydrolase by kaempferol and related naturally occurring flavonoids. *British Journal of Pharmacology*, 155(2):244–252, 2009.

Tohge, Takayuki; Nishiyama, Yasutaka; Hirai, Masami Yokota; Yano, Mitsuru; Nakajima, Jun-ichiro; Awazuhara, Motoko; Inoue, Eri; Takahashi, Hideki; Goodenowe, Dayan B; Kitayama, Masahiko, and others, . Functional genomics by integrated analysis of metabolome and transcriptome of arabidopsis plants over-expressing an MYB transcription factor. *The Plant Journal*, 42(2):218–235, 2005.

Toufighi, Kiana; Brady, Siobhan M; Austin, Ryan; Ly, Eugene, and Provart, Nicholas J. The botany array resource: E-northerns, expression angling, and promoter analyses. *The Plant Journal*, 43(1):153–163, 2005.

Tucker, Matthew R; Okada, Takashi; Hu, Yingkao; Scholefield, Andrew; Taylor, Jennifer M, and Koltunow, Anna MG. Somatic small RNA pathways promote the mitotic events of megagametogenesis during female reproductive development in arabidopsis. *Development*, 139(8):1399–1404, 2012.

Tung, Chih-Wei; Dwyer, Kathleen G; Nasrallah, Mikhail E, and Nasrallah, June B. Genome-wide identification of genes expressed in arabidopsis pistils specifically along the path of pollen tube growth. *Plant Physiology*, 138(2):977–989, 2005.

Updegraff, Emily P; Zhao, Fang, and Preuss, Daphne. The extracellular lipase EXL4 is required for efficient hydration of arabidopsis pollen. *Sexual Plant Reproduction*, 22(3):197–204, 2009.

Usami, Takeshi; Horiguchi, Gorou; Yano, Satoshi, and Tsukaya, Hirokazu. The more and smaller cells mutants of *Arabidopsis thaliana* identify novel roles for SQUAMOSA PROMOTER BINDING PROTEIN-LIKE genes in the control of heteroblasty. *Development*, 136(6):955–964, 2009.

Van Damme, Mireille; Huibers, Robin P; Elberse, Joyce, and Van den Ackerveken, Guido. Arabidopsis DMR6 encodes a putative 2og-fe (ii) oxygenase that is defense-associated but required for susceptibility to downy mildew. *The Plant Journal*, 54(5):785–793, 2008.

Van Hoek, Milan and Hogeweg, Paulien. The effect of stochasticity on the lac operon: an evolutionary perspective. *PLoS Computational Biology*, 3(6):e111, 2007.

Van Leene, Jelle; Hollunder, Jens; Eeckhout, Dominique; Persiau, Geert; Van De Slijke, Eveline; Stals, Hilde; Van Isterdael, Gert; Verkest, Aurine; Neirynck, Sandy; Buffel, Yelle, and others, . Targeted interactomics reveals a complex core cell cycle machinery in arabidopsis thaliana. *Molecular Systems Biology*, 6(1), 2010.

Varbanova, Marina; Yamaguchi, Shinjiro; Yang, Yue; McKelvey, Katherine; Hanada, Atsushi; Borochov, Roy; Yu, Fei; Jikumaru, Yusuke; Ross, Jeannine; Cortes, Diego, and others, . Methylation of gibberellins by arabidopsis GAMT1 and GAMT2. *The Plant Cell Online*, 19(1):32–45, 2007.

Wahl, Vanessa; Brand, Luise H; Guo, Ya-Long, and Schmid, Markus. The FANTASTIC FOUR proteins influence shoot meristem size in *Arabidopsis thaliana*. *BMC Plant Biology*, 10(1):285, 2010.

Wang, Congmao; Marshall, Alex; Zhang, Dabing, and Wilson, Zoe A. ANAP: An integrated knowledge base for arabidopsis protein interaction network analysis. *Plant Physiology*, 158(4):1523–1533, 2012.

Wang, Guo-Ying; Shi, Jiang-Li; Ng, Gina; Battle, Stephanie L; Zhang, Chong, and Lu, Hua. Circadian clock-regulated phosphate transporter PHT4; 1 plays an important role in arabidopsis defense. *Molecular plant*, 4 (3):516–526, 2011.

Wang, Jia-Wei; Czech, Benjamin, and Weigel, Detlef. mir156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. *Cell*, 138(4):738–749, 2009.

Wang, Mingyi; Benedito, Vagner Augusto; Zhao, Patrick Xuechun, and Udvardi, Michael. Inferring large-scale gene regulatory networks using a low-order constraint-based algorithm. *Molecular BioSystems*, 6(6):988–998, 2010.

Weichert, Annett; Brinkmann, Christopher; Komarova, Nataliya Y; Dietrich, Daniela; Thor, Kathrin; Meier, Stefan; Suter Grotemeyer, Marianne, and Rentsch, Doris. AtPTR4 and atPTR6 are differentially expressed, tonoplast-localized members of the peptide transporter/nitrate transporter 1 (PTR/NRT1) family. *Planta*, 235(2):311–323, 2012.

Wellmer, Frank; Alves-Ferreira, Márcio; Dubois, Annick; Riechmann, José Luis, and Meyerowitz, Elliot M. Genome-wide analysis of gene expression during early arabidopsis flower development. *PLoS Genetics*, 2(7): e117, 2006.

Wigge, Philip A; Kim, Min Chul; Jaeger, Katja E; Busch, Wolfgang; Schmid, Markus; Lohmann, Jan U, and Weigel, Detlef. Integration of spatial and temporal information during floral induction in arabidopsis. *Science Signalling*, 309(5737):1056, 2005.

Wijeratne, Asela J; Zhang, Wei; Sun, Yujin; Liu, Wenlei; Albert, Reka; Zheng, Zhengui; Oppenheimer, David G; Zhao, Dazhong, and Ma, Hong. Differential gene expression in arabidopsis wild-type and mutant anthers: insights into anther cell differentiation and regulatory networks. *The Plant Journal*, 52(1):14–29, 2007.

Wilson, Iain W; Kennedy, Gavin C; Peacock, James W, and Dennis, Elizabeth S. Microarray analysis reveals vegetative molecular phenotypes of arabidopsis flowering-time mutants. *Plant and Cell Physiology*, 46(8): 1190–1201, 2005.

Wimburly, Frank C; Heiman, Thomas; Ramsey, Joseph, and Glymour, Clark. Experiments on the accuracy of algorithms for inferring the structure of genetic regulatory networks from microarray expression levels. *Proc.* IJCAI *2003 Bioinformatics Workshop*, 2003.

Winkel-Shirley, Brenda. Flavonoid biosynthesis. a colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiology*, 126(2):485–493, 2001.

Winkel-Shirley, Brenda. Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways. *Physiologia Plantarum*, 107(1):142–149, 2002.

Wynn, April N; Rueschhoff, Elizabeth E, and Franks, Robert G. Transcriptomic characterization of a synergistic genetic interaction during carpel margin meristem development in *Arabidopsis thaliana*. *PloS One*, 6(10): e26231, 2011.

Xiang, Daoquan; Yang, Hui; Venglat, Prakash; Cao, Yongguo; Wen, Rui; Ren, Maozhi; Stone, Sandra; Wang, Edwin; Wang, Hong; Xiao, Wei, and others, . POPCORN functions in the auxin pathway to regulate embryonic body plan and meristem organization in arabidopsis. *The Plant Cell Online*, 23(12):4348–4367, 2011.

Yamaguchi, Ayako; Wu, Miin-Feng; Yang, Li; Wu, Gang; Poethig, R Scott, and Wagner, Doris. The microrna-regulated SBP-box transcription factor SPL3 is a direct upstream activator of *LEAFY*, *FRUITFULL*, and *APETALA1*. *Developmental Cell*, 17(2):268–278, 2009.

Yamanishi, Yoshihiro; Vert, Jean-Philippe, and Kanehisa, Minoru. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(suppl 1):i468–i477, 2005.

Yamasaki, Hideo; Sakihama, Yasuko, and Ikehara, Norikatsu. Flavonoid-peroxidase reaction as a detoxification mechanism of plant cells against H2O2. *Plant Physiology*, 115(4):1405–1412, 1997.

Yang, Caiyun; Vizcay-Barrena, Gema; Conner, Katie, and Wilson, Zoe A. MALE STERILITY1 is required for tapetal development and pollen wall biosynthesis. *The Plant Cell Online*, 19(11):3530–3548, 2007.

Yang, Fang; Wang, Quan; Schmitz, Gregor; Müller, Dörte, and Theres, Klaus. The bHLH protein ROX acts in concert with RAX1 and LAS to modulate axillary meristem formation in arabidopsis. *The Plant Journal*, 2012.

Yang, Yue; Xu, Richard; Ma, Choong-je; Vlot, A Corina; Klessig, Daniel F, and Pichersky, Eran. Inactive methyl indole-3-acetic acid ester can be hydrolyzed and activated by several esterases belonging to the atMES esterase family of arabidopsis. *Plant Physiology*, 147(3):1034–1045, 2008.

Yant, Levi; Mathieu, Johannes; Dinh, Thanh Theresa; Ott, Felix; Lanz, Christa; Wollmann, Heike; Chen, Xuemei, and Schmid, Markus. Orchestration of the floral transition and floral development in arabidopsis by the bifunctional transcription factor APETALA2. *The Plant Cell Online*, 22(7):2156–2170, 2010.

Yonekura-Sakakibara, Keiko; Tohge, Takayuki; Matsuda, Fumio; Nakabayashi, Ryo; Takayama, Hiromitsu; Niida, Rie; Watanabe-Takahashi, Akiko; Inoue, Eri, and Saito, Kazuki. Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene–metabolite correlations in arabidopsis. *The Plant Cell Online*, 20(8):2160–2176, 2008.

Yoo, Changwon; Thorsson, Vesteinn; Cooper, Gregory F, and others, . Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. In *Proceedings of Pacific Symposium on Biocomputing*, volume 7, pages 498–509, 2002.

Yoshimoto, Hiroyuki; Saltsman, Kirstie; Gasch, Audrey P; Li, Hong Xia; Ogawa, Nobuo; Botstein, David; Brown, Patrick O, and Cyert, Martha S. Genome-wide analysis of gene expression regulated by the calcineurin/crz1p signaling pathway in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, 277(34):31079–31088, 2002.

Yu, Hee-Ju; Hogan, Pat, and Sundaresan, Venkatesan. Analysis of the female gametophyte transcriptome of arabidopsis by comparative expression profiling. *Plant Physiology*, 139(4):1853–1869, 2005.

Yu, Jing; Smith, V Anne; Wang, Paul P; Hartemink, Alexander J, and Jarvis, Erich D. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18): 3594–3603, 2004.

Yu, Qing-Bo; Li, Guang; Wang, Guan; Sun, Jing-Chun; Wang, Peng-Cheng; Wang, Chen; Mi, Hua-Ling; Ma, Wei-Min; Cui, Jian; Cui, Yong-Lan, and others, . Construction of a chloroplast protein interaction network and functional mining of photosynthetic proteins in *Arabidopsis thaliana*. *Cell Research*, 18(10):1007–1019, 2008.

Zhang, Huiming; Kim, Mi-Seong; Krishnamachari, Venkat; Payton, Paxton; Sun, Yan; Grimson, Mark; Farag, Mohamed A; Ryu, Choong-Min; Allen, Randy; Melo, Itamar S, and others, . Rhizobacterial volatile emissions regulate auxin homeostasis and cell expansion in arabidopsis. *Planta*, 226(4):839–851, 2007.

Zhang, Jiji and Spirtes, Peter. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18 (2):239–271, 2008.

Zhang, Xiaohong; Feng, Baomin; Zhang, Qing; Zhang, Diya; Altman, Naomi, and Ma, Hong. Genome-wide expression profiling and identification of gene activities during early flower development in arabidopsis. *Plant Molecular Biology*, 58(3):401–419, 2005.

Zhang, Xiujun; Zhao, Xing-Ming; He, Kun; Lu, Le; Cao, Yongwei; Liu, Jingdong; Hao, Jin-Kao; Liu, Zhi-Ping, and Chen, Luonan. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1):98–104, 2012.

Zhao, Dazhong; Ni, Weimin; Feng, Baomin; Han, Tianfu; Petrasek, Megan G, and Ma, Hong. Members of the arabidopsis-SKP1-like gene family exhibit a variety of expression patterns and may play diverse roles in arabidopsis. *Plant Physiology*, 133(1):203–217, 2003.

Zhao, Jinfeng; Zhang, Wenhui; Zhao, Yang; Gong, Ximing; Guo, Lei; Zhu, Guoli; Wang, Xuechen; Gong, Zhizhong; Schumaker, Karen S, and Guo, Yan. SAD2, an importin β-like protein, is required for UV-B response in arabidopsis by mediating MYB4 nuclear trafficking. *The Plant Cell Online*, 19(11):3805–3818, 2007.

Zhou, Aifen and Li, Jia. Arabidopsis BRS1 is a secreted and active serine carboxypeptidase. *Journal of Biological Chemistry*, 280(42):35554–35561, 2005.

Zhu, Huifen; Qian, Weiqiang; Lu, Xuzhong; Li, Dongping; Liu, Xin; Liu, Kunfan, and Wang, Daowen. Expression patterns of purple acid phosphatase genes in arabidopsis organs and functional analysis of atPAP23 predominantly transcribed in flower. *Plant Molecular Biology*, 59(4):581–594, 2005.

Zik, Moriyah and Irish, Vivian F. Global identification of target genes regulated by APETALA3 and PISTILLATA floral homeotic gene action. *The Plant Cell Online*, 15(1):207–222, 2003.

Zuber, Hélène; Davidian, Jean-Claude; Aubert, Grégoire; Aimé, Delphine; Belghazi, Maya; Lugan, Raphaël; Heintz, Dimitri; Wirtz, Markus; Hell, Rüdiger; Thompson, Richard, and others, . The seed composition of arabidopsis mutants for the group 3 sulfate transporters indicates a role in sulfate translocation within developing seeds. *Plant Physiology*, 154(2):913–926, 2010.